MASTER'S THESIS

De politiek-strategische governance van artificial intelligence bij militaire toepassingen

Derks, M. (Mike)

Award date:
2021

De politiek-strategische governance van artificial intelligence bij militaire toepassingen

The Political Strategic governance of artificial intelligence in military applications

Opleiding:	Open Universiteit, faculteit Management, Science & Technology
	Masteropleiding Business Process Management & IT
Programme:	Open University of the Netherlands, faculty of Management, Science & Technology Master Business Process Management & IT
Cursus:	IMA0001 Afstuderen Business Process Management and IT IM9806 Afstudeeropdracht Business Process Management and IT
Student:	drs. M.C.E.J. Derks bsc.
Identiteitsnummer:	
Datum:	14 september 2021
Afstudeerbegeleider	dr. L. Bollen
Meelezer	prof.dr.ir. R. Helms
Versie nummer:	2.1
Status:	definitief


Manipulatie
Oversight
Diversiteit
Betrouwbaar
Menselijk
Interactie
Leren
Schadelijk
Security
Begrip
Eerlijk
Afhankelijk
Verslaving
Privacy
Toegankelijkheid
Logging
Verslaglegging
Robuust
Weerbaar
Onafhankelijk
Milieu

Abstract

Technologie is een belangrijke factor die de combat effectiveness van een leger bepaald. Daarbij wordt Artificial Intelligence gezien als een van de technologieën die het aanzien van defensie gaat veranderen. Maar wat is Artificial intelligence? Hoe gaan we de ontwikkeling en inzet beheersen? Dat start bij de top, de Tweede Kamer. Deze paper behandeld de Governance van Artificial Intelligence vanuit het Trustworthy AI Governance model. Op basis van een wordcount is een match gemaakt tussen de documenten van de Tweede Kamer en het Trustworthy AI Governance model. Daarbij komt naar voren dat de focus van de Tweede Kamer meer ligt op Transparantie, data governance en privacy. Elementen als milieu, sociale omgeving, diversiteit en non-discrimination komen veel minder vaak voor.

Sleutelbegrippen

Trustworthy Governance, Militaire toepassingen, Politiek-strategisch, Artificial Intelligence, Autonomous Weapon Systems

Samenvatting

Technologie wordt vaak gezien als een meerwaarde voor het optreden van militaire eenheden. Het hebben van technologische hulpmiddelen geeft de soldaat een voorsprong op zijn tegenstander. Technologische voorsprong vergroot de slagkracht van een leger. Artificial Intellingence (AI) wordt gezien als één van de belangrijkste ontwikkelingen voor het leger van de toekomst.

De governance van AI is een vakgebied in ontwikkeling. Verschillende modellen worden ontwikkeld maar nog geen enkel model is ontwikkeld op het gebied voor militaire toepassingen.

Binnen de governance structuur van defensie ligt het mandaat voor ontwikkeling en inzet van geweldsmiddelen bij het politiek-strategisch niveau. Doel van deze paper is te komen tot een model wat gebruikt kan worden om governance uit te voeren op specifieke investeringen en inzetten van AI in het militaire domein. Daarbij is als centrale probleemstelling geformuleerd:

> Op welke manier kan politiek-strategische governance uitgeoefend worden bij het gebruik van AI in militaire toepassingen?

Om deze vraag te beantwoorden zijn een aantal stappen genomen. Allereerst is gekeken naar de vraag: Wat is AI. AI kan geclassificeerd worden vanuit vijf verschillende invalshoeken. De onderkende invalshoeken zijn: Rationaliteit, mate van impact, mate van intelligentie, dimensies en militaire toepassingsgebieden. De vijf dimensies geven daarbij de verschillende verschijningsvormen van AI weer.

Daarnaast is een theoretisch model gebouwd. Daarbij is gebruik gemaakt van het Trustworthy model. Dit model is ontwikkeld door een commissie van de Europeese Unie en gebaseerd op de ethische/mensenrechten uit de "charter of fundamental rights and international human right law". In dit model komen een zevental kernwaarden terug waaraan een AI model moet voldoen, te weten #1 Human Agency and Oversight, #2 Technical Robustness and Safety, #3 Privacy and Data governance, #4 Transparency, #5 Diversity, Non-discrimination, and Fairness, #6 Societal and Environmental Well-being, #7 Accountability.

De keuze voor het model van de Trustworthy AI Governance is gedaan op basis van de aanname dat governance op politiek-strategisch niveau over een tweetal zaken zal gaan: de rol van de overheid (regelgevende en monitorende instantie) en de nadruk op de mogelijke menselijke "schade" die op kan treden (ethische vraagstukken).

Daarnaast is een empirisch onderzoek gedaan naar de relatie tussen het model Trustworthy AI Governance en de documenten van de Tweede Kamer. Daartoe is gebruik gemaakt van een zogenaamde "wordcount". De documenten van de Tweede Kamer zijn geanalyseerd op woordgebruik en deze woordlijst is gematched met een woordlijst uit het Trustworthy AI Governance model. Hiertoe zijn de documenten uit de database van de Tweede Kamer verzameld, geconverteerd, verwerkt en uiteindelijk gematched.

In het empirisch onderzoek komt naar voren dat alle kernwaarde uit het model voor komen, maar dat de verdeling niet gelijkmatig is. De kernwaarden #3 Privancy and Data governance (22%), #4 Transparency (28%) en #6 Societal and Environmental Well-being (16%) komen vaker voor, terwijl kernwaarden als #5 Diversity, Non-discrimination, and Fairness (4%) en#7 Accountability (8%) minder vaak in documenten voor komen.

Als je de analyse twee niveaus dieper maakt, zie je dat er kernwoorden zijn die meer en kernwoorden die minder gebuikt worden. Woorden als "toezicht", "privacy", "integriteit", "risico" en "milieu" zijn veel voorkomende woorden. Er wordt dus meer over de elementen #1.2 Human Oversight, #3 Privacy and Data governance, #4 Transparency, #4.3 Communication en #6.1 Environmental Well-being gesproken.

Conclusie hieruit zou zijn dat, waar het model uit gaat van redelijk gelijkmatige verdeling, de focus in de praktijk naar een beperkt aantal kernwaarden zal gaan. Politieke voorkeuren, specifieke situationele gebeurtenissen of modelmatige onvolkomenheden zouden hier de oorzaak van kunnen zijn, maar dit is niet uit de data af te leiden.

In de huidige politiek-strategische governance worden bepaalde kernwoorden/-begrippen/-waarden vaker gebruikt. De focus ligt meer op data governance en privacy. Mogelijk zijn dat, op dat moment, de elementen waarover gesproken moet worden. Maar het risico bestaat ook dat er elementen vergeten worden. Voor de praktijk ligt hier een verantwoordelijkheid om de volledigheid te bewaken. Alle elementen die een rol spelen in het model zullen bewaakt moeten worden, waarbij de juiste balans gezocht wordt.

Het Trustworthy AI model is relatief nieuw. Er zijn nog geen empirische onderzoeken naar gedaan. Dit onderzoek geeft een eerste empirische toetsing van het model, echter binnen een specifiek vakgebied (militaire toepassingen), binnen een specifiek werkveld (politiek-strategisch) en binnen een beperkte periode (2011-2021). Om de rigiditeit en de juistheid van het model te toetsen zal aanvullend onderzoek noodzakelijk zijn.

Summary

Technology is one of the added values for a soldier in the battlefield. Having access to technology gives the soldier an edge over his opponent. Technology increases the battle capabilities of armies. Artificial Intelligence (AI) is one of the major developments for the army of the future.

Governing AI development and use is an upcoming field. Several models have been developed, but nonspecific for the use on military applications.

Approving the development and the application of force is mandated to the political-strategic level. This paper aims to develop a model that can be used to govern the development and application of AI in military applications. Central lies the following question:

> In what way can political-strategic governance be executed in the use of military AI application.

To answer this quenstion, one should first answer the question: What is AI. Basically, AI can be classified using 5 perspectives, being rationality, impact, intelligence, dimensions, and military use of AI.

The theoretical model used in this paper is built on the Trustworthy model. This model was developed by the European Union Commission and based on the ethics and human rights from the charter of fundamental rights and international human right law. This model recognizes seven core values for AI development: #1 Human Agency and Oversight, #2 Technical Robustness and Safety, #3 Privacy and Data governance, #4 Transparency, #5 Diversity, Non-discrimination, and Fairness, #6 Societal and Environmental Well-being, and #7 Accountability.

The Trustworthy AI Governance was selected on the premise that governance on a political level is directed by two main issues: The role of the government as regulating and monitoring and the possible (negative) effects that can occur.

Empirical research was done into the relation between the Trustworthy AI Governance model and the Tweede Kamer. Using the word-count method, documents from the Tweede Kamer were analyzed and matched to the words used in the model. For this purpose, documents from the Tweede Kamer were collected, converted, processed, and eventually matched.

The results show that the distribution among the 7 classes is not evenly. The elements #3 Privancy and Data governance (22%), #4 Transparency (28%) and #6 Societal and Environmental Well-being (16%) had a higher percentage than #5 Diversity, Non-discrimination, and Fairness (4%) and #7 Accountability (8%).

On a deeper level, certain words were used more frequently. “toezicht”, “privacy”, “integriteit”, “risico” and “milieu” were frequently used, leading to high values in the classes #1.2 Human Oversight, #3 Privacy and Data governance, #4 Transparency, #4.3 Communication and #6.1 Environmental Well-being

When looking at the model one would expect an evenly distribution, but the empirical data does not support this. Political preference, specific events of errors in the model might be the cause. But no conclusion to these factors can be made on the available data.

The strategic political governance is focuses on transparency, data governance and privacy. Other elements were also included, but to a lesser extent. This poses the risk that specific elements/classes are unjustly excluded from the discussion. All elements should be addressed. The users of the model should guard against excluding elements.

Trustworthy AI Governance is relatively new. No empirical research had been done to the model. This might be the first research into this field. It is within the field of Military Applications. But this is also the drawback of the research. The limited and very specific field, combined with the limited time. Additional research is required to complete and fully test the model

Inhoudsopgave

Lijst met afbeeldingen

Lijst met tabellen

1. Introductie

1.1. Achtergrond

Technologie wordt vaak gezien als een meerwaarde voor het optreden van militaire eenheden. Het hebben van technologische hulpmiddelen geeft de soldaat een voorsprong op zijn tegenstander. Technologische voorsprong vergroot de slagkracht van een leger. Maar technologie is niet de enige factor van belang. Uiteindelijk zijn ook kennis, vaardigheden, procedures, morele overtuiging en aantallen van belang. De geschiedenis laat zien dat technologie niet de enige factor is. In de tweede wereldoorlog behaalde het Duitse leger grote winsten door technologische voorsprong (en goede procedures), maar verloren uiteindelijk op de aantallen. Technologische voorsprong is een factor van belang, maar niet de enige.

Het tot je beschikking hebben van een technologie is niet altijd onomstreden. In de middeleeuwen werd het gebruik van de kruisboog als dusdanig afgrijselijk en godslasterlijk gezien dat Paus Urbanus II het gebruik van de kruisboog in 1097 verbood bij conflicten tussen Christenen onderling. Met het inzetten tegen niet Christenen had hij kennelijk minder moeite. De inzet van de kruisboog was niet onomstreden. Uiteindelijk werd de kruisboog een onderdeel van de bewapening van de militairen.

Ook bij het moderne optreden van defensie wordt nieuwe technologie als een belangrijke driver gezien. In de defensie doctrine (Defensiestaf, 2019) wordt conceptuele innovatie met daarin de implementatie van eigen ontwikkelde, maar ook commercieel ontwikkelde, nieuwe technologieën onderkend. Daarnaast wordt naast de "traditionele" maritieme, land- en luchtdomeinen ook nog het ruimtedomein, het cyberspace domein, het elektromagnetische spectrum en het akoestische spectrum onderkend.

Eén van de technologieën die in opkomst is, is artificial intelligence (AI). Door het Belfer Center for Science and International Affairs, onderdeel van de HARVARD Kennedy School, wordt AI gezien als de technologie die in de toekomst de doorslag zal geven. Dit omdat AI zowel van invloed is op het fysieke (militaire) domein als op het informatie en het economische domein.(Allen & Chan, 2017). Het hebben van AI-middelen wordt gezien als doorslaggevend bij militaire inzet.

"Artificial intelligence is the future, not only for Russia, but for all humankind. It comes with colossal opportunities, but also threats that are difficult to predict. Whoever becomes the leader in this sphere will become the ruler of the world," zei de Russische president Vladimir Putin tijdens een open les. China, USA en Rusland hebben het onderwerp AI hoog op de militaire agenda staan. Europese landen volgen op een afstand. Ook Nederland heeft AI op de agenda staan. AI wordt gezien als één van de belangrijkste ontwikkelingen op defensiegebied.

Op dit moment is AI vooral nog "mens" gedreven. De AI-toepassing geeft advies, maar het uiteindelijke besluit wordt door de mens genomen. Er zijn echter geen beperkingen die de ontwikkeling en inzet van "Lethal Autonomous Weapon Systems" verbieden, maar binnen het Amerikaanse leger bijvoorbeeld worden deze nog niet gebruikt. (Hoadley & Sayler, 2019) De eindbeslissing om tot actie over te gaan ligt nog altijd bij de mens.

Maar in hoeverre is de mens nog in staat om het advies van het AI te begrijpen. De input van de AI komt soms uit bronnen die voor een mens niet zonder vertaling begrijpelijk zijn (bijvoorbeeld elektromagnetische signalen). Daarnaast is, in een aantal gevallen, de verwerkingssnelheid van de AI vele malen hoger dan het menselijk vermogen. Maar ook de gebruikte algoritmen veranderen onder invloed van de omgeving. Dit maakt het erg lastig om de uitkomst van de AI te begrijpen. De operator moet bijna blind vertrouwen op de uitkomst.

Voordat het zo ver is, moet eerst het politieke besluit genomen worden om te investeren in AI-toepassingen. Zoals al eerder aangegeven is niet iedere technologische investering onomstreden en ook AI is daar geen uitzondering op. Of er nu gebruik gemaakt wordt van commercieel

ontwikkelde toepassingen of van zelf ontwikkelde toepassingen, er zal in geïnvesteerd moeten worden en daarvoor zijn middelen benodigd. Middelen die hiervoor vrijgemaakt moeten worden of gealloceerd moeten worden. En zodra er middelen beschikbaar gemaakt moeten worden, zijn er altijd politieke en economische overwegingen of dit wel of niet gedaan moet worden.

De politieke overwegingen vormen de basis voor het wel of niet ontwikkelen en inzetten van AI gestuurde middelen.

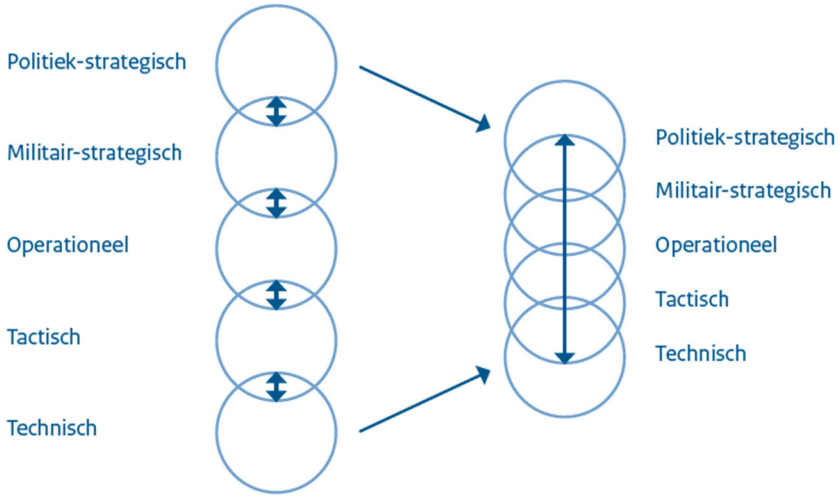
1.2. Gebiedsverkenning

De governance van AI is een vakgebied in ontwikkeling. Verschillende modellen worden ontwikkeld (Cotter, 2015; Gasser & Almeida, 2017; Haven, 2019; Siau & Wang, 2018), maar nog geen enkel model is ontwikkeld op het gebied voor militaire toepassingen. De modellen die ontwikkeld worden zijn vooral voor gebruik in de civiele markt en de overstap naar militaire toepassingen wordt in een aantal gevallen als ethisch bezwaarlijk gezien. Er zijn gevallen bekend waarin het bedrijf uit ethische overwegingen weigerde om met defensie samen te werken (Hoadley & Sayler, 2019). Daaruit rijst de vraag in hoeverre de bestaande modellen voor governance bruikbaar zijn in het militaire domein.

Zoals eerder aangegeven ligt de beslissing voor het ontwikkelen en inzetten van AI-middelen bij de politiek. Echter de politiek staat daarbij niet alleen, maar hangt samen met verschillende niveaus in de defensie kolom. Daarbij worden de volgende niveaus onderkend:

- Politiek-strategisch;
- Militair-strategisch;
- Operationeel;
- Tactisch en
- Technisch.

Binnen de governance structuur ligt het mandaat voor ontwikkeling en inzet van geweldsmiddelen bij het politiek-strategisch niveau. De vaststelling van de Grand Strategy is ook de (exclusieve) verantwoordelijkheid van het politiek-strategische niveau. Daarnaast formuleert zij aanvullende richtlijnen voor het gebruik van de machtsmiddelen, zonder daarbij in detail te treden.



Figuur 1: Strategische compressie

Door toenemende technische mogelijkheden en de real-time informatievoorziening bestaat de mogelijkheid voor het politiek-strategisch niveau om de uitvoering door het technisch niveau in detail te volgen. Ook in omgekeerde richting kan handelen op bijvoorbeeld tactisch niveau directe invloed hebben voor het politiek-strategisch niveau. Dit wordt Strategische Compressie genoemd. (Defensiestaf, 2019).

Strategische compressie zorgt er voor dat informatie en gegevens snel tussen de niveaus gedeeld worden, maar de verantwoordelijkheid voor het gebruik van geweldsmiddelen ligt nog steeds bij het politiek-strategische niveau. Datzelfde geld ook voor de investering. In de begroting van defensie zijn de investeringen opgenomen in het defensie investeringsplan (DIP) onder verantwoordelijkheid van de Minister van Defensie. Governance hierop wordt uitgevoerd door de Tweede Kamer (vertegenwoordigd door de vaste kamer commissie defensie).

1.3. Probleemstelling en deelvragen

Zoals aangegeven liggen besluiten over investeringen en inzet van machtsmiddelen exclusief bij het politiek-strategische niveau. De politiek beslist welke middelen worden ontwikkeld en of deze worden ingezet. Dat geld ook voor een onderwerp als AI-toepassingen en Lethal Autonomous Weapons. Maar op basis van welke motieven nemen de politici deze besluiten? Wat beweegt hen

om een investering wel of niet goed te keuren? En welke zakelijke of emotionele overwegingen maken ze dan?

Als je weet hoe mensen tot een besluit komen, kun je een inhoudelijke discussie voeren over het ontwikkelen en inzetten van machtsmiddelen als AI. Je weet dan over welke factoren er verantwoording afgelegd moet worden en hoe belangrijk die factoren zijn.

Doel van deze paper is te komen tot een model wat gebruikt kan worden om governance uit te voeren op specifieke investeringen en inzetten van AI in het militaire domein. Daarbij is als centrale probleemstelling geformuleerd:

> Op welke manier kan politiek-strategische governance uitgeoefend worden bij het gebruik van AI in militaire toepassingen?

1.4. Opdrachtformulering

De verdere uitwerking van de probleemstelling zal gedaan worden aan de hand van de volgende deelvragen:

1. Wat is AI?
2. Hoe worden AI vertaald naar een militaire context?
   Wat zijn de nu gebruikte definities van AI in de literatuur en hoe kunnen deze samengebracht worden tot één praktische definitie in een militaire context?

3. Hoe kan governance van AI worden ingericht?
4. Hoe vertaalt de governance zich naar kernwaarden voor AI binnen een militaire context?
   Welke modellen worden er in de theorie nu gebruikt voor de governance van AI en welk model is toepasbaar in een militaire politieke situatie?

5. Welk belang geeft "de politiek" aan de verschillende meetdimensies?
   Empirisch onderzoek naar de mate waarin de verschillende dimensies van belang zijn voor governance van militaire toepassingen.

1.5. Motivatie/relevantie

Door de strategische compressie (zie paragraaf 1.2) komen steeds meer niveaus met elkaar in contact. Daarbij zullen militairen die tactisch bezig zijn soms een andere "taal" spreken dan de politici van het strategisch-politieke niveau. Dit probleem wordt nog verder versterkt door verschillende operationele commando's die verschillende taken hebben en verschillende middelen gebruiken. Het vakjargon van de verschillende disciplines binnen defensie kennen elk hun eigen woorden en termen. Hierdoor is de communicatie tussen de medewerkers soms lastig en wordt er veel tijd en energie gestoken in een goede samenwerking.

Maar de politiek laat zich in deze minder gemakkelijk sturen. Iedere vier jaar (of minder) zijn er verkiezingen en kan het politiek-strategisch niveau deels wisselen en treden er nieuwe politici aan. Hierdoor kan het voor militairen lastig zijn om met politici te communiceren. Nieuwe politici kunnen anders tegen de wereld aan kijken en andere kernwaarden belangrijk vinden.

De literatuur onderzoekt de governance generiek. Zo wordt bijvoorbeeld de ISO/IEC 38507 uitgewerkt voor de governance van AI, van een generieke visie. Er wordt in de modellen niet specifiek gekeken naar militaire toepassingen.

Hierdoor ontstaat een drieluik wat niet met elkaar in balans is. Theorie, politiek en werkvloer kunnen verschillende standpunten en visies hebben.

In dit onderzoek wordt een theoretisch model van governance vertaald naar een militaire toepassing. Die vervolgens getoetst wordt aan de politieke praktijk om de militaire "werkvloer" de middelen te geven om over de juiste kernbegrippen met het politiek-strategische niveau te communiceren. Hierdoor zal de afstemming in de toekomst kunnen verbeteren.

1.6. Leeswijzer

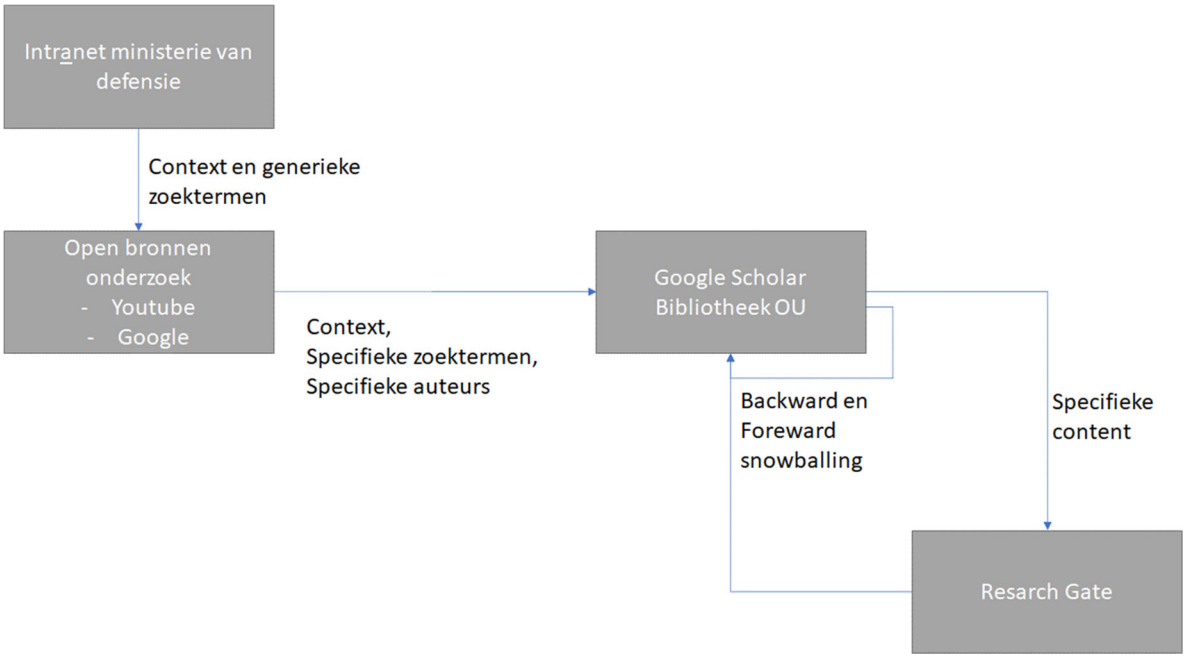
In de volgende hoofdstukken worden achtereenvolgens de volgende elementen behandeld. In hoofdstuk 2 zal een theoretische beschouwing gegeven worden van AI en governance daarvan. Vervolgens wordt in hoofdstuk 3 het methodologische kader behandeld om het empirische onderzoek van governance te doen. In hoofdstuk 4 zal het resultaat van het daadwerkelijke onderzoek worden weergegeven. De thesis sluit af met de conclusies en beschouwing van de samenhang van het geheel in hoofdstuk 5.

2. Theoretisch kader

Het eerste deel van het onderzoek is het theoretische kader. Daarin worden vooral de grenzen van het onderzoek aangegeven. Dat bestaat uit een aantal elementen. Allereerst zal paragraaf 2.1 ingaan op de onderzoekaanpak, hoe is het (theoretische) onderzoek tot stand gekomen. Daarna wordt in paragraaf 2.2 weergegeven hoe het onderzoek naar de literatuur is uitgevoerd, welke bronnen zijn gebruikt en wat de resultaten waren. In paragraaf 2.3 en 2.4 zal dan het daadwerkelijke theoretische kader inhoudelijk vorm worden gegeven. Daarbij wordt per deelvraagstuk aangegeven wat de uitkomsten zijn. Tenslotte sluit het hoofdstuk af in paragraaf 2.5 met de conclusies en vervolgstappen.

2.1. Onderzoeksaanpak

De aanpak van het theoretische onderzoek is gelaagd. Daarbij wordt breed begonnen en per stap verder verfijnd. Vanuit de gevonden resultaten wordt met backward en foreward snowballing het theoretisch kader verdiept. Om de wetenschappelijke kwaliteit zo veel mogelijk te bewaken is gebruik gemaakt van verschillende bronnen. Dit is grafisch weergegeven in Figuur 2.



Figuur 2: Grafische weergave zoekproces

1e zoekslag Intranet

De zoektocht is gestart op het Ministerie van Defensie Internet Netwerk (intranet). De zoeksleutel "artificial intelligence" leverde een beperkt aantal hits op. Het kenniscentrum voor innovatie defensie heeft in 2018 artificial intelligence geïdentificeerd als één van de kernontwikkelingen voor de toekomst. Vandaar dat er een sharepoint is ingericht met een verzameling van documenten. Deze documenten zijn gebruikt om via de backward snowball methode een tweetal documenten te identificeren die als basis gebruikt kunnen worden (Allen & Chan, 2017; Hoadley & Sayler, 2019)

2e zoekslag Internet/Youtube

De tweede zoekslag is gemaakt via internet/youtube. Doelstelling van deze zoektocht was niet zozeer te zoeken naar literatuur als naar toelichting op modellen en personen die zich in het veld met deze materie bezighouden. Gebruikte sleutel "artificial intelligence governance" levert een aantal hits op. Door sortering naar relevantie, staan volgens de Google algoritme relevantste verwijzingen vooraan. Dit levert een tweetal interessante video's op (Dafoe, 2019; Haven, 2019). Wetenschappelijk gezien is deze methode van sortering echter twijfelachtig omdat persoonlijke voorkeur van de onderzoeker in het algoritme wordt meegenomen. Echter gezien de resultaten heb ik toch besloten om deze methode te gebruiken. De inhoud van Youtube veranderd zo snel dat iedere sorteersleutel binnen zeer korte tijd niet meer te herhalen is.

3e zoekslag Bibliotheek OU/Google Scollar/Researchgate

De derde ronde van de zoekslag heeft plaatsgevonden in de OU online bibliotheek, Google scollar en Researchgate. Hierbij gebruik makend van zowel de foreward als de backward snowball methode. Daarbij zijn ook specifieke zoektermen gebruikt die aansluiten bij de deelvragen, in tegenstelling tot de eerste twee slagen die generieker van aard waren. In onderstaande tabel zijn de resultaten samengevat weergegeven:

Tabel 1: Samenvatting literatuuronderzoek deelvragen 1-4

Deelvraag	Gevonden Referenties (aantal)	Bekeken Referenties (aantal)	Gebruikte Referenties (aantal)
1. Wat is AI? 2. Hoe worden AI vertaald naar een militaire context?	9412	663	25
3. Hoe kan governance van AI worden ingericht? 4. Hoe vertaalt de governance zich naar kernwaarden voor AI binnen een militaire context?	1511	241	9

De details van het onderzoek zijn opgenomen in Tabel 10: Overzicht zoekresultaten literatuuronderzoek in detail deelvragen 1-4 in BIJLAGE I

2.2. Classificatie artificial intelligence

Als je in de literatuur zoekt naar de vraag wat is AI, kom je veel literatuur tegen. AI is een container begrip. Verschillende definities worden door elkaar gebruikt. Om actieve governance uit te kunnen oefenen moet eerst duidelijk zijn wat het precies is en wat niet. In deze paragraaf ga ik in op de subvragen

1. Wat is AI?
2. Hoe wordt AI vertaald in een militaire context?

2.2.1. Situationele definitie artificial intelligence

De eerste deelvraag is een zoektocht naar de generieke definitie van AI. Er is geen generieke definitie van AI. In de loop van de tijd zijn verschillende definities in verschillende situaties gebruikt. Hierdoor ontstaan tijd- en situatiegebonden definities. De definitie is afhankelijk van het specifieke veld waar de onderzoeker in zit en de technologische ontwikkelingen van die tijd.

Om echter te begrijpen wat AI is, is het goed om te begrijpen hoe de definitie in de loop van de tijd veranderd is en hoe verschillende situaties tot verschillende definities leiden. Verschillende invalshoeken spelen daarbij een rol.

Hieronder een overzicht van verschillende manieren van definitie naar rationaliteit, impact en intelligentie.

➢ *Rationaliteit*

Naarmate technologische ontwikkelingen plaats vinden, ontstaan er verschillende visies op AI (Hoadley & Sayler, 2019). Daarbij ontwikkeld de definitie van AI zich van menselijk denken en handelen naar rationeel denken en handelen. In tabel 2 zijn deze definities in de loop van de tijd weergegeven

Tabel 2: Definities van AI in de tijd

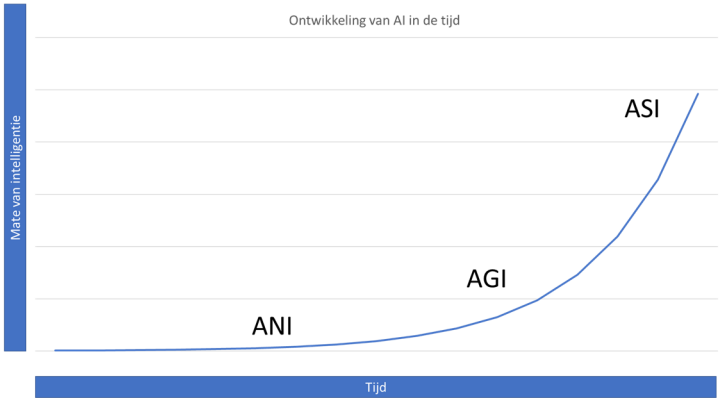
1978 Systemen die denken als mensen: *"The automation of activities that we associate with human thinking, activities such as decision making, problem solving, and learning."*	1992 Systemen die rationeel kunnen denken: *"The study of computations that make possible to perceive, reason, and act."*
1990 Systemen die handelen als mensen: *"The art of creating machines that perform functions that require intelligence when performed by people."*	1993 Systemen die rationeel kunnen handelen: *"The branch of computer science that is concerned with the automation of intelligent behavior."*

Daarbij zien we in de loop van de tijd verschillende dimensies ontstaan, van menselijk denken naar het rationeel handelen.

➢ *Mate van impact*

De mate van impact is een andere factor die in de definitie een rol speelt. Daarbij ontwikkeld de definitie zich van een smalle naar een brede visie.

De huidige eerste generatie van AI, de zogenaamde artificial narrow intelligence (ANI), is vooral gericht op het verrichten van eenvoudige taken. Het herkennen van gezichten of gesproken commando's verwerken. In de toekomst zal begrip en redenering aan de AI worden toegevoegd. De AI wordt dan in staat autonoom een taak te vervullen of een probleem op te lossen. In de theorie wordt dit Artificial General Intelligence genoemd (AGI). Mogelijk komt er in de toekomst nog een derde vorm van AI die zelfbewust is. In de theorie heet dit de Artificial Super Intelligence (ASI). Deze is dan volledig zelfstandig en zou de mens overbodig maken. Het is deze vorm van AI waar Elon Musk[1] en Bill Gates[2] in de reguliere media voor waarschuwen als de ondergang van de mensheid.



Figuur 3: Ontwikkeling van AI in de loop van de tijd

[1] https://www.theguardian.com/technology/2014/oct/27/elon-musk-artificial-intelligence-ai-biggest-existential-threat

[2] https://www.washingtonpost.com/news/the-switch/wp/2015/01/28/bill-gates-on-dangers-of-artificial-intelligence-dont-understand-why-some-people-are-not-concerned/

➢ *Mate van intelligentie*

De derde factor die van belang is binnen de classificatie van AI is de mate van intelligentie van het systeem:

Kaplan & Haenlein maken een onderscheid in cognitieve[3], emotionele[4] en sociale[5] intelligentie (Kaplan & Haenlein, 2019). Op basis van de mate waarin deze vormen van intelligentie aanwezig zijn worden een drietal vormen van AI onderkend:

- Analytical AI: Dit is op dit moment de meest voorkomende vorm van AI. Bijvoorbeeld fraude detectie of de zoekmachine van Google. Deze AI-toepassing heeft alleen een cognitieve intelligentie. Door middel van extrapolatie van het verleden wordt een beeld ontwikkeld van wat er in de toekomst waarschijnlijk zal gebeuren. De AI gebruikt dit als referentie van handelen.
- Human-Inspired AI: In de Human-Inspired AI komen zowel elementen voor van cognitieve intelligentie als emotionele intelligentie. Het gaat dan niet alleen om het herkennen van trends, maar ook om het herkennen van emotie. De handelingen van de AI zijn niet alleen afhankelijk van de extrapolatie uit het verleden, maar ook gebaseerd op menselijke emoties en andere non-verbale signalen.
- Humanized AI: Deze systemen bestaan nog niet. De AI-systemen hebben zowel cognitieve, als emotionele en sociale intelligentie. Daarbij zijn de AI-systemen zelfbewust en in staat om zelfstandig te handelen.
  (Hoadley & Sayler, 2019)

2.2.2. Artificial intelligence in militaire context

Al met al bestaat er in de theorie geen specifieke eenduidige definitie van AI. Ook als je de vertaalslag naar de militaire context gaat maken loop je tegen dezelfde beperkingen en problemen aan. Vandaar dat je in onderzoeksrapporten over AI altijd ziet dat ze beginnen met het aangeven van een definitie. (Allen & Chan, 2017; Bogdanoski & Nacev; Grega et al., 2019; Hoadley & Sayler, 2019; Manen, Sweijs, & Arkhipov-Goyal, 2019). Daarbij verschillen de definities naar gelang de doelstelling van het rapport, of het specifieke gebied waarnaar men kijkt.

Het Amerikaanse Departement of Defence (DOD) heeft, ten behoeve van het congres, geprobeerd een overkoepelende definitie te maken van AI. Deze staat in de National Defense Authorization Act(Hoadley & Sayler, 2019):

1. "Any artificial system that performs tasks under varying and unpredictable circumstances without significant human oversight, or that can learn from experience and improve performance when exposed to data sets.
2. An artificial system developed in computer software, physical hardware, or other context that solves tasks requiring human-like perception, cognition, planning, learning, communication, or physical action.
3. An artificial system designed to think or act like a human, including cognitive architectures and neural networks.
4. A set of techniques, including machine learning that is designed to approximate a cognitive task.
5. An artificial system designed to act rationally, including an intelligent software agent or embodied robot that achieves goals using perception, planning, reasoning, learning, communicating, decision-making, and acting."

Alhoewel deze definitie zeer omvangrijk is, is hij ook gedreven door de huidige toepassingen zoals die voor AI onderkend worden. In de verdere uitwerking van de mogelijkheden en de daarvan

[3] "e.g. competencies related to pattern recognition and systematic thinking" (Kaplan & Haenlein, 2019)
[4] "e.g. adaptability, selfconfidence, emotional self-awareness, achievement orientation" (Kaplan & Haenlein, 2019)
[5] "e.g. empathy, teamwork, inspirational leadership" (Kaplan & Haenlein, 2019)

afgeleide AI-strategie heeft de DOD de definitie meer losgelaten en is gaan kijken naar de toepassingsgebieden die AI beslaan. Daarbij onderkend men de volgende gebieden (Corn, 2019):

1. Verbeteren van situational awareness en besluitvorming. Denk daarbij aan image recognition in het kader van intel verzamelen. Maar ook de verwerking van grote hoeveelheden ruwe data tot bruikbare informatie.
2. Verbeteren van de veiligheid van materieel. Denk daarbij aan sensoren die de veiligheid van de operator verbeteren
3. Voorspellen/vaststellen van onderhouds- en bevoorradingsmomenten.
4. Verbeteren van bedrijfsprocessen.

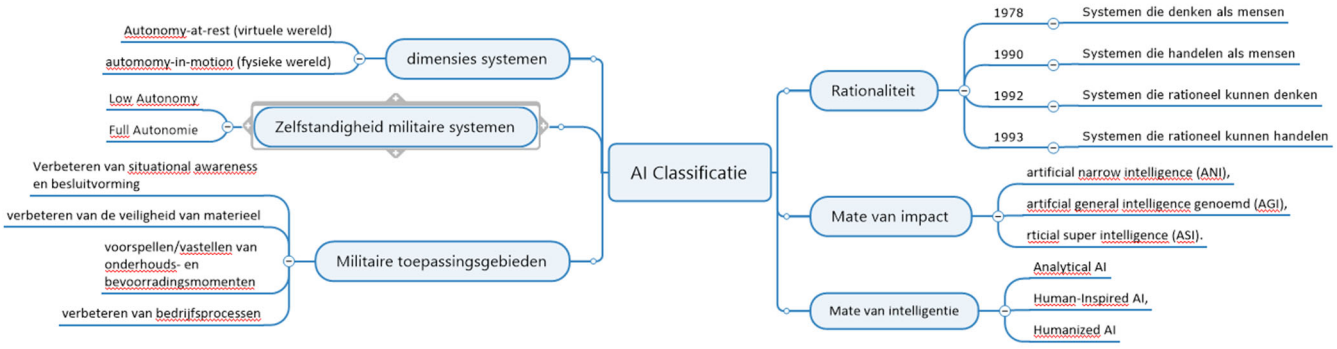
Ook Morgan e.a. hebben een classificatie gemaakt van militaire toepassingen. (Morgan et al., 2020) Daarbij maken zij een indeling naar zelfstandigheid en dimensie van operatie.

Binnen de zelfstandigheid maken ze onderscheid naar mate van zelfstandigheid in handelen. Daarbij onderscheiden ze aan de ene kant van het spectrum "low autonomy" ook wel geautomatiseerde taken genoemd en in het andere uiterste "fully autonomous".

Daarnaast maken ze ook een onderscheid naar de dimensie waarin AI zich bevindt. "Autonomy-at-rest" wordt gebruikt voor toepassing van AI in de virtuele wereld en aan de andere kant onderkennen ze "automomy-in-motion" wat zich hoofzakelijk in de fysieke wereld af speelt. De meest extreme vorm hiervan zijn "Lethal Autonomous Weapon Systems", de zogeheten LAWS. Wapensystemen die, na afvuren, volledig zelfstandig kunnen opereren.

2.2.3. Conclusies definitie

Zowel vanuit de generieke kant als vanuit de specifieke militaire kant is geen eenduidige definitie van AI te geven. Gegeven definities veranderen in de loop van de tijd of zijn context gedreven. In de generieke literatuur wordt gekeken naar Rationaliteit, Mate van Impact en Mate van intelligentie. Terwijl in de militair gerichte literatuur gekeken wordt naar dimensies (fysiek of virtueel), Zelfstandigheid en Toepassingsgebieden. In Figuur 4 is dit schematisch weergegeven.



Figuur 4: Classificatie AI

2.3. Model van governance

Nadat we in de vorige paragraaf een theoretische verkenning gemaakt hebben van de deelvragen 1 en 2, de definitie van AI, zowel generiek als specifiek militair. Wordt in deze paragraaf verder ingegaan op de deelvragen 3 en 4:

3. Hoe kan governance van AI worden ingericht?
4. Hoe vertaalt de governance zich naar kernwaarden voor AI binnen een militaire context?

2.3.1. Verschillende modellen governance

Er zijn geen governance modellen die specifiek kijken naar de governance van AI binnen de militaire context. Ook in hun artikel "Artificial intelligence Regulation: A Meta-Frameword for Formulation and Governance" (Almeida, Santos, & Farias, 2020) onderkennen Almeida e.a.

verschillende modellen, maar geen van deze modellen is specifiek gericht op de militaire toepassingen.

Om toch een selectie van een model te kunnen maken is een aanname gedaan. Politiek-strategische governance van AI bij militaire toepassingen vindt plaats binnen de verantwoordelijkheid van de Tweede Kamer. Aanname hierbij is dat de nadruk ligt op de rol van de overheid (regelgevende en monitorende instantie) en de mogelijke menselijke "schade" die op kan treden (ethische vraagstukken).

Op basis van deze aanname is een shortlist gemaakt van modellen die hierop zouden kunnen passen. Dit waren de volgende modellen: Competency-Based AI, Sustainable AI Development, en Trusworthy AI development. Deze modellen zijn meer in de diepte bekeken. Op basis van de bevindingen hieruit is het model van Trustworthy AI perspectief geselecteerd omdat deze het meeste bij de gestelde aanname aansluit.

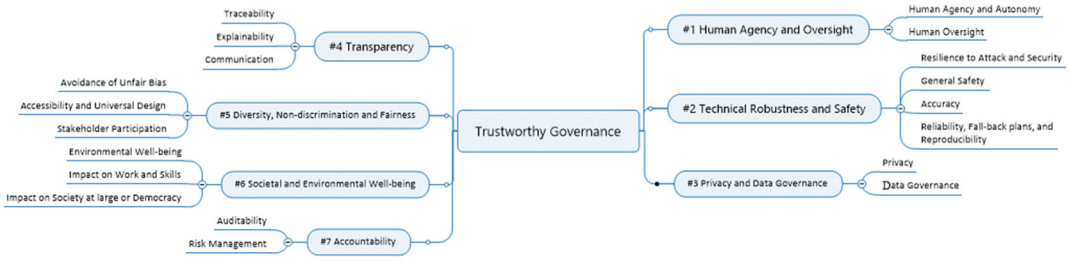
2.3.2. Governance vanuit het Trustworthy perspectief

De basis voor Trustworthy AI Governance is het ethische/mensenrechten vraagstuk. Trustworthy AI gaat volgens de High level expert group van de Europeese Commissie (EU) uit van systemen die rechtmatig, ethisch en robuust zijn (High-Level_Expert_Group_on_Artificial_Intelligence, 2020).

Het model van de EU gaat uit van wat ze noemen de "fundamentele rechten". Dit zijn de volgende 4 rechten:

1. "Does the AI system potentially negatively discriminate against people on the basis of any of the following grounds (non-exhaustively): sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation?"
2. "Does the AI system respect the rights of the child, for example with respect to child protection and taking the child's best interests into account?"
3. "Does the AI system protect personal data relating to individuals in line with GDPR?"
4. "Does the AI system respect the freedom of expression and information and/or freedom of assembly and association?"

Deze rechten worden dan vertaald in 7 kernwaarden waaraan systemen moeten voldoen. Per kernwaarde worden kernbegrippen geformuleerd. Vervolgens worden de kernbegrippen weer vertaald in vragen. Met behulp van de vragen kan de gebruiker inzicht krijgen of aan de kernwaarde is voldaan. In figuur 5 zijn de 7 voorwaarden met bijbehorende kernbegrippen grafisch weergegeven.



Figuur 5: Trustworthy AI Governance

2.3.3. Governance vanuit militair perspectief

Defensie is een organisatie die geweldsmiddelen kan en mag inzetten. Dit geeft een bijzondere positie. In wetgevingen en verdragen is de reikwijdte van deze inzet geregeld. Per inzet wordt bepaald onder welke wettelijke regels de inzet valt. Dat kan zijn in de vorm van Standard Operating Procedures (SOP) of missie-specifieke Rules of Engagement (ROE). De governance op deze inzet is in de grondwet geregeld. Artikel 100 van de grondwet legt deze taak bij de Staten-Generaal. Echter in dezelfde grondwet zijn ook afwijkingen weergegeven bijvoorbeeld bij de noodtoestand of staat van oorlog.

De basis van de militaire governance ligt in een aantal specifieke wetten en verdragen. Deze zijn uitgewerkt in het militair oorlogsrecht in een drietal uitgangspunten (Voetelink, 2012). Deze uitgangspunten zijn:

A. Het recht van soevereine staten om geweld te gebruiken binnen het collectief veiligheidssysteem van de Verenigde Naties, het zogeheten *ius ad bellum*;
B. Beperking en bescherming van slachtoffers van oorlogsgeweld, het zogeheten *humanitair oorlogsrecht*;
C. Waarborgen van de mensenrechten, zowel voor, tijdens als na het conflict.

Om deze regels meer handen en voeten te geven zijn er een drietal principes geformuleerd waar iedere inzet aan wordt gemeten (Ducheine, 2008):

1) Noodzakelijkheidsvereiste – De inzet van dit geweldsmiddel is noodzakelijk om het doel te bereiken;
2) Proportionaliteit – Het gebruikte geweldsmiddel staat in verhouding tot het doel;
3) Subsidiariteit – Er is geen lichter middel beschikbaar om dit doel te bereiken.

Daarmee is governance binnen het militaire domein vooral een juridisch gedreven vraagstuk geworden.

2.4. Conclusie

In dit hoofdstuk is een theoretisch model opgebouwd. Daarbij is het model opgebouwd vanuit een breed profiel naar een steeds smaller wordend profiel. Uiteindelijk zijn een viertal deelvragen beantwoord.

Allereerst is gekeken naar de vraag: Wat is AI. Daarbij is vanuit een generiek model de doorvertaling gemaakt naar een specifiek militaire context. Vanuit vijf verschillende invalshoeken wordt het onderwerp AI geclassificeerd. De onderkende invalshoeken zijn: Rationaliteit, mate van impact, mate van intelligentie, dimensies en militaire toepassingsgebieden.

De tweede deelvraag die beantwoord is, is de vraag van governance op AI binnen militaire toepassingen. Ook daarbij is de analyse gemaakt van generiek naar specifiek. In het generieke model voor governance is gekozen voor het Trustworthy model. Dit model is gemaakt door een commissie van de Europeese Unie en gebaseerd op de ethische/mensenrechten uit de "charter of fundamental rights and international human right law".

De keuze voor het model van de Trustworthy AI Governance is gedaan op basis van de aanname dat governance op politiek-strategisch niveau over een tweetal zaken zal gaan: de rol van de overheid (regelgevende en monitorende instantie) en de nadruk op de mogelijke menselijke "schade" die op kan treden (ethische vraagstukken).

De militaire basis voor governance ligt vooral in wet- en regelgeving. Vanuit de grondwet waarbij de governance van militaire inzet is belegd bij de Staten-Generaal, naar vertalingen van afspraken binnen de United Nations in wet- en regelgeving. Hiervan is het militair Oorlogsrecht de overkoepelende wetgeving.

In het Militair Oorlogsrecht zijn een drietal uitgangspunten verweven: Soevereiniteit van de staten, bescherming slachtoffers en waarborgen mensenrechten. Deze zijn uiteindelijk vertaald in drie principes: Noodzakelijkheidsvereiste, proportionaliteitsvereiste en subsidiariteitsvereiste.

Met de formulering van een theoretisch model is het onderzoek niet afgerond. De vraag kan gesteld worden in hoeverre het theoretische model in de specifieke Nederlandse praktijk ook zo werkt. In de deelvragen is dit deelvraag 5:

5. Welk belang geeft "de politiek" aan de verschillende meetdimensies?

Hieruit kan de vervolgvraag geformuleerd worden:

A. Wordt de governance door de Nederlandse Politiek uitgevoerd met gebruikmaking van de kernwaarden zoals weergegeven in het Trustworthy AI Governance model?

Met andere woorden, komen de elementen zoals die in het Trustworthy AI Governance Model voorkomen ook terug in de feitelijk Nederlandse Politieke governance. En komen ze in gelijke mate voor of wordt er meer nadruk gelegd op specifieke deelgebieden.

De selectie van het Trustworthy AI Governance model is gedaan op basis van een aanname. Hieruit kan een subvraag worden afgeleid:

B. Is de aanname dat de Nederlandse politiek-strategische governance vooral plaats vindt op basis van de kernelementen wetgeving/monitoring en ethisch/mensenrechten juist?

Als de aanname juist is, zou uit de analyse van deelvraag A moeten blijken dat de wetgeving/monitoring en ethisch/mensenrechten gerelateerde elementen van het model evenredig of vaker voorkomen dan de anderen.

3. Methodologie

3.1. Inleiding

Hoofdstuk 2 is geëindigd met een aantal vervolgvragen die nodig zijn om de centrale vraagstelling te beantwoorden:

> Op welke manier kan politiek-strategische governance uitgeoefend worden bij het gebruik van AI in militaire toepassingen?

Na de theoretische verkenning komen we nu aan bij de praktische uitwerking van de vraag. Hoe wordt de governance van AI bij militaire toepassingen in Nederland uitgevoerd?

Dit hoofdstuk zal ingaan op de methodologische onderbouwing van het empirische deel van het onderzoek. Allereerst zullen in paragraaf 3.2 de keuzes zoals gemaakt worden toegelicht. In paragraaf 3.3 wordt het technisch ontwerp toegelicht en in paragraaf 3.4 wordt duidelijk gemaakt hoe de gegevens worden geanalyseerd. In paragraaf 3.5 zal nog een reflectie gegeven worden op betrouwbaarheid, de validiteit en ethische aspecten. Tenslotte sluit paragraaf 3.6 af met de conclusie.

3.2. Keuzes in het onderzoek

Het empirische deel van het onderzoek bestaat uit een deductief onderzoek. De beschreven theorie van de Trustworthy AI Governance wordt getoetst aan de specifieke Nederlandse politiek-strategische situatie. Dit wordt gedaan door middel van een Archiefonderzoek. In dit archiefonderzoek worden de documenten (verslagen, notulen, agenda's etc..) van de Tweede Kamer getoetst aan het model om de correlatie tussen beiden vast te stellen. Uiteindelijk moet dat resulteren in een correlatie overzicht van de verschillende steekwoorden die in het model voorkomen, gerelateerd aan de documenten van de Tweede Kamer.

De keuze voor deductieve onderzoeksstrategie is gebaseerd op het feit dat er sprake is van de toetsing van een theoretisch model. In tegenstelling tot de inductieve methode, waarbij theorie wordt ontwikkeld, wordt hier het model getoetst en op basis van de resultaten eventueel aangepast.

De gekozen methode van onderzoek is een archiefonderzoek. In een archiefonderzoek worden secundaire databronnen in een systeem geladen en geschikt gemaakt voor analyse. In dit specifieke geval worden documenten van de Tweede Kamer geladen en geanalyseerd op woordgebruik. Op deze wijze wordt een kwantitatieve analyse van de tekstuele verslagen mogelijk en kan de correlatie tussen het model en de theorie worden vastgesteld.

Door gebruik te maken van de open data van de Tweede Kamer kan een inzicht in de correlatie gegeven worden, gerelateerd aan de tijd. Daarnaast zijn de documenten van de Tweede Kamer onderdeel van een officieel gepubliceerde dataset, de historische verslaglegging. Dit maakt het mogelijk om het onderzoek later te herhalen of uit te breiden.

Er zijn een drietal methoden mogelijk om de data te benaderen:

> Methode #1: Maken van een script waarmee de data vanuit de open databronnen gegenereerd kan worden om deze daarna te analyseren
>
> Methode #2: Handmatig de data uit de zoekmachine van de 2e kamer halen om een dataset op te bouwen die gebruikt kan worden om te analyseren.
>
> Methode #3: Zelf een dataset opbouwen in een analyseomgeving en deze gebruiken voor het onderzoek.

De voorkeur in het onderzoek zou uitgaan naar methode #1. In deze methode wordt de onderzoekdata uit de officieel gepubliceerde documenten gehaald met behulp van een script. Echter op dit moment is het niet mogelijk om toegang tot de dataset te krijgen omdat er geen nieuwe gebruikers worden toegelaten. Om de set te optimaliseren en toegang te vereenvoudigen wordt de database opnieuw opgebouwd. Deze herbouw zit ten tijde van het onderzoek in de BETA

testen en de verwachting is niet dat dit op korte termijn beschikbaar zal zijn. Vandaar dat deze methode af valt.

Methode #2 is het handmatig opvragen van alle documenten op basis van de zoekmachine. Daarmee wordt de dataset doorzocht op dezelfde wijze als met methode #1, alleen handmatig. Dit zou tot dezelfde resultaten moeten leiden als het script. Door het grote aantal mogelijke zoektermen, is dit een zeer arbeidsintensieve methode. 18 dimensies in de classificatie * 46 zoektermen in 2 talen (Nederlands en Engels), levert minimaal 1656 handmatige opvragen op. Alhoewel dit technisch haalbaar is, is er daarna geen mogelijkheid meer om de data verder te analyseren of vervolgonderzoeken te doen. Vandaar dat gekozen is om deze methode niet te gebruiken.

Uit praktische overwegingen, zoals aangegeven in de vorige alinea, wordt in het onderzoek gebruik gemaakt van methode #3, het zelf opbouwen van een dataset. Door de documenten in een eigen analyseomgeving te zetten en daarna te converteren naar een analyseerbare datastructuur, kunnen verschillende analyses gemaakt worden. Daarnaast is de dataset dan nog beschikbaar voor verdiepingsonderzoek.

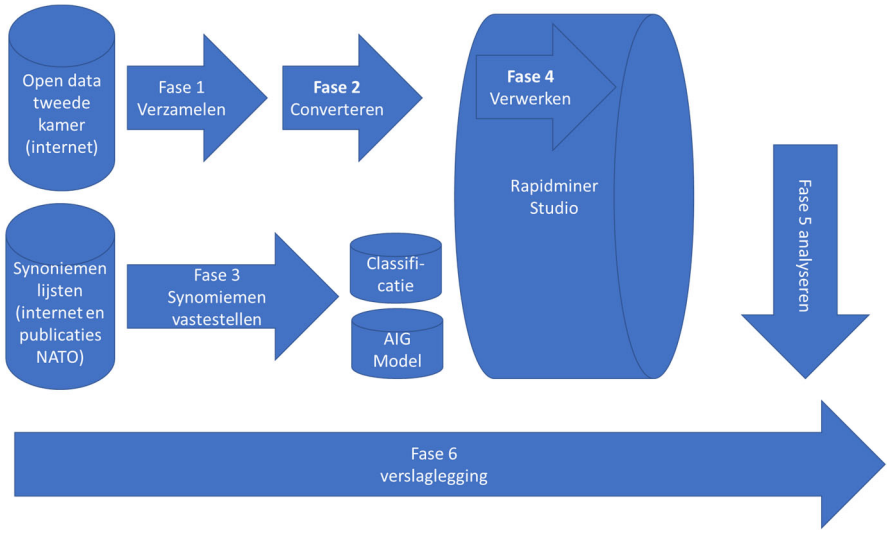
3.3. Content Analyse

De methode van content analyse kan zowel een kwantitatieve als kwalitatieve analyse zijn. "Quantitative content analysis is a deductive approach that tests research hypotheses after systematically coding data into variables, and qualitative content analysis is an inductive method that reaches conclusions after an open and in-depth analysis of texts" (Nefes, 2020). Daarbij wordt de scheiding tussen kwalitatieve en kwantitatieve analyse vrij strikt toegepast. Vanwege het grote aantal documenten wat in de analyse beschikbaar is, is gekozen de kwantitatieve methode te volgen, de zogeheten dictionary methode (Benoit, 2014) of "Wordscore" (Laver, Benoit, & Garry, 2003).

In deze methode worden documenten omgezet in woordenlijsten/-frequenties, waarna deze output gebruikt kan worden voor het toetsen van bestaande modellen. Door deze analyse wordt het Trustworthy AI Governance model getoetst aan de wordlist. Daarbij vindt de toetsing van de referentiewoorden op het laagste niveau plaats en worden deze daarna geaggregeerd.

3.4. Technisch ontwerp: uitwerking van de methode

De uitwerking van het empirische deel van het onderzoek zal plaatsvinden in een vijftal fases. Deze fases zijn deels parallel uitgevoerd en deels sequentieel. In figuur 6 zijn de verschillende fases in relatie tot elkaar en de data weergegeven.



Figuur 6: Fases in het analyse proces

Fase 1: Verzamelen van de brondata

In deze fase worden de documenten van de 2e kamer verzameld/gedownload uit de publieke database. Dit betreft alle documenten (die niet geclassificeerd zijn als geheim) met betrekking tot de Tweede Kamer. Om niet alle documenten te downloaden (dat zou te omvangrijk zijn), is de selectie "defensie" gebruikt, binnen de eerder gestelde periode van 10 jaar (maart 2011-juni 2021). In

totaal zijn er op die manier 21.505 documenten succesvol gedownload[6]. Omdat batch download niet mogelijk was, is dit allemaal handmatig gedaan op 1684 webpagina's met totaal ruim 23.000 muisklikken.

Fase 2: Converteren

Om de documenten te kunnen verwerken, zijn de documenten geconverteerd naar TXT-bestanden. Hiervoor is de tool Doxillion Document Converter® gebruikt. Hierdoor konden de bestanden enigszins automatisch verwerkt worden. Alleen door de grote omvang van het aantal documenten moesten batches van ca 500-2000 stukt opgestart worden. De verwerking is daarbij goed gegaan en de PDF en WORD documenten zijn verwerkt. 57 documenten konden niet verwerkt worden en moesten handmatig geconverteerd worden. 10 documenten konden niet verwerkt worden. Dit omdat de documenten geen tekst bevatten, maar bijvoorbeeld uitsluitend afbeeldingen.

Fase 3: Bouw referentiemodel woorden

Om de analyse op basis van de classificatie en het governance model uit te kunnen voeren moeten beide modellen omgezet worden in referentiewoorden.

Voor de Engelse termen is een Nederlands synoniem vastgesteld. Bijna alle documenten binnen de Staten-Generaal zijn opgesteld in de Nederlandse taal[7]. Daarbij worden wel specifieke Engelstalige vaktermen gebruikt, maar ook Nederlandse synoniemen. Om de juiste termen vast te stellen zijn alle vaktermen/steekwoorden via Google translate gecontroleerd.

Fase 4: Verwerking

Om de daadwerkelijke match tussen de woordenlijst en het referentiemodel te kunnen maken, zijn de geselecteerde documenten omgezet in een woordenlijst. Daarvoor is een programma in Rapidminer® geschreven om de documenten te selecteren, converteren in een wordlist en te verwerken in verwerkbare output. De detailuitwerking hiervan is terug te vinden in BIJLAGE III.

Fase 5: Uitvoeren van de analyse

De analyse die uitgevoerd is, is een linguïstische analyse. Specifieke methode die daarbij gebruikt wordt in de kwantitatieve tekstanalyse op basis van wordscore. Daarbij wordt het woordgebruik in documenten bekeken door ze te matchen met een kernwoordenlijst. Uit technische overwegingen is dit gedaan in Excel.

Fase 6: Verslaglegging

De verslaglegging (dit rapport) heeft plaatsgevonden parallel aan de uitvoering. Dit om de transfer tussen observatie en vastlegging zo kort mogelijk te houden.

3.5. Reflectie t.a.v. betrouwbaarheid, validiteit en ethische aspecten

De betrouwbaarheid van de methode wordscore is hoog (Laver et al., 2003). Daarvoor zijn een aantal redenen:

- De methode van wordscore is een technisch eenvoudige methode.
- De methode is herhaalbaar voor andere/latere onderzoeken. Verslagen van de Staten Generaal en commissies zijn openbaar en vormen als zodanig een permante (toetsbare) bron die gebruikt kan worden;

[6] Amendementen (36) Besluitenlijsten (1113) Brieven regering (3773) Commissieverslagen (1057) Kamervragen (2380) Moties (655) Overige Kamerstukken (11399) Plenaire verslagen (734) Stemmingsuitslagen (149) Verslagen (157) Wetsvoorstellen (52)

[7] Bij de verwerking is 1 document gevonden waarvan de bijlage uitsluitend in het Engels was (Macro Implications of Micro Transformations: An Assessment of AI's Impact on Contemporary Geopolitics (Manen et al., 2019)). Discussie over dit document was echter volledig in het Nederland

- De onderzoeker kan geen invloed uitoefenen (direct of indirect) op de inhoud van de data;
- De conclusies die uit de data komen en eventuele verdiepingsvragen, kunnen op dezelfde dataset worden gedaan. Zonder dat, naar verwachting, aanvullende data verzameld hoeft te worden.

Daarentegen zijn er wel een aantal zaken die de betrouwbaarheid negatief beïnvloeden:

- Er zijn een aantal conversies nodig om tot verwerkbare bestanden te komen. Gedurende de conversies gaat data verloren. Welke kan achteraf niet, of met zeer veel moeite, vastgesteld worden;
- De conversies worden handmatig uitgevoerd. Alhoewel dat met de grootst mogelijke zorg en met controlegetallen gebeurt, kunnen fouten gemaakt worden;
- Daarnaast zijn er nog een aantal specifieke aandachtspunten voor de teksten zoals gesteld door Laver e.a.:
  i) Teksten moet opgesteld zijn in gelijkwaardige lexicografische niveaus.
  ii) De gebruikte teksten moeten betrekking hebben op het politieke onderwerp wat bestudeerd wordt.
  iii) De teksten moeten zo veel mogelijk verschillende woorden bevatten.

Validiteit is een ander punt van zorg in de methode van de wordscore (Benoit, 2014). In een document kunnen verschillende woorden gebruikt worden, zonder dat er een correlatie tussen de beide woorden bestaat. De kans op een onjuiste constatering is hierdoor aanwezig.

Echter, de validiteit van de analyse zit verankerd in de functie van de Tweede Kamer. Toezicht houden op de werking van de ministeries is de hoofdtaak van de Tweede Kamer[8] en analyses hierover zouden dan mogelijk niet betrekking hebben op het specifieke onderwerp, maar wel op de generieke taakstelling.

Tenslotte nog een opmerking over de ethische implicatie van het doen van een wordscore-onderzoek. De analyse van de documenten van de Tweede Kamer gaat over reacties en opmerkingen die Tweede Kamer leden hebben gemaakt gedurende vergaderingen. Het gebruik van deze documenten zou de integriteit van deze personen kunnen schaden. Dit is echter niet het geval vanwege de volgende redenen:

- De documenten vallen onder de Wet Openbaarheid Bestuur (WOB) en alle documenten die gepubliceerd zijn, zijn onderdeel van de openbare “geschiedschrijving”;
- Documenten die niet onder de WOB vallen, bijvoorbeeld vanwege de classificatie “geheim” zijn niet gepubliceerd en kunnen daarom niet gebruikt worden;
- Er wordt niet specifiek gekeken naar individuele personen die deel uitmaken van de Tweede Kamer, maar naar de algemene woorden die gebruikt zijn. Daarbij is de relatie tussen de persoon en het gebruikte woord niet meer aanwezig.

Vandaar dat de documenten zonder beperking gebruikt kunnen worden, zonder dat daarmee individuele personen geschaad worden.

8 https://www.tweedekamer.nl/zo_werkt_de_kamer/de_nederlandse_democratie/taken_en_rechten

4. Empirische resultaten

4.1. Inleiding

Nadat in de vorige hoofstukken gekeken is naar het theoretisch en methodologisch kader, zullen in dit hoofdstuk de resultaten worden weergegeven. Dit hoofdstuk bestaat voor een groot deel uit cijfers en tabellen, dat is de output van het onderzoek. Per tabel wordt aangegeven wat de eerste constateringen zijn n.a.v. de data.

De opbouw van de data is gefaseerd. Allereerst wordt het hoogste niveau besproken om daarna verder de diepte in te gaan. Daarbij worden niet alle details besproken. Die worden weergegeven in de BIJLAGE IV.

Allereerst wordt in paragraaf 4.2 een overzicht van de output gegeven zoals die gevonden is. In paragraaf 4.3 wordt ingegaan op de gevoeligheidsanalyse. Tenslotte wordt in paragraaf 4.4 het totaaloverzicht gegeven van de analyse, na de correcties die uit de gevoeligheidsanalyse komen. Het hoofdstuk sluit af met een conclusie in paragraaf 4.5

4.2. Output

De output van het model is opgebouwd uit een aantal verschillende niveaus. Daarbij is het theoretische Trustworthy AI Governance model zoals beschreven in hoofdstuk 2 de basis. De in het model gebruikte niveaus zijn zo ook in de analyse gebruikt. Daarbij zijn de volgende niveaus onderkend:

Niveau 1: De 7 kernwaarden (weergegeven als N1)

Niveau 2: De 7 kernwaarden zijn onderverdeeld in een aantal kernbegrippen (N2)

Niveau 3: De kernbegrippen zijn weer vertaald in Nederlandse kernwoorden. (N3)

Voor een compleet overzicht van de vertaling van kernwaarden naar kernbegrippen naar kernwoorden verwijs ik naar BIJLAGE II.

4.2.1. Niveau 1 output:

Zoals aangegeven is het hoogste niveau in het model de 7 kernwaarden. In onderstaande tabel staat per kernwaarde aangegeven in hoeveel documenten woorden voorkomen die betrekking hebben op deze kernwaarde. Dat is de telling van de documenten waarin woorden, of combinaties van woorden, één of meerdere keren voor komen.

Tabel 3: N1 resultaten

Kernwaarde (N1)	Aantal documenten
#1 Human Agency and Oversight	83
#2 Technical Robustness and Safety	46
#3 Privacy and Data governance	95
#4 Transparency	284
#5 Diversity, Non-discrimination, and Fairness	18
#6 Societal and Environmental Well-being	68
#7 Accountability	36

Wat hierin meteen opvalt is dat de verdeling over de kenwaarden niet gelijkmatig is. #4 Transparency komt veel vaker voor dan #5 Diversity, Non-discrimination, and Fairness (284 resp. 18). Deze scheve verdeling heeft te maken met extreme scores op een aantal waarden in de details. De kernwoorden “kwaliteit”, “communicatie” en “risico” komen binnen de kernwaarde #4 Transparancy erg vaak voor. Dit zijn redelijk generieke woorden.

4.2.2. Niveau 2 output

In onderstaande tabel staat per kernwaarde de onderverdeling naar kernbegrippen aangegeven en het aantal documenten waarin woorden voorkomen die betrekking hebben op deze kernbegrippen.

Tabel 4: N2 resultaten

Kernwaarden (N1)	Kernbegrip (N2)	Aantal documenten
#1 Human Agency and Oversight	#1.1 Human Agency and Autonomy	18
	#1.2 Human Oversight	65
#2 Technical Robustness and Safety	#2.1 Resilience to Attack and Security	31
	#2.2 General Safety	8
	#2.3 Accuracy	31
	#2.4 Reliability, Fall-back plans, and Reproducibility	65
#3 Privacy and Data governance	#3.1 Privacy	60
	#3.2 Data governance	35
#4 Transparency	#4.1 Traceability	77
	#4.2 Explainability	7
	#4.3 Communication	200
#5 Diversity, Non-discrimination, and Fairness	#5.1 Avoidance of Unfair Bias	14
	#5.2 Accessibility and Universal Design	4
	#5.3 Stakeholder Participation	0
#6 Societal and Environmental Well-being	#6.1 Environmental Well-being	60
	#6.2 Impact on Work and Skills	3
	#6.3 Impact on Society at large or Democracy	5
#7 Accountability	#7.1 Auditability	2
	#7.2 Risk Management	34

Het beeld wat in Niveau 1 te zien is, dat #4 Transparency erg hoog scoort, is ook in de tabel Niveau 2 te zien. #4.1 Traceability wordt volledig bepaald door het kernbegrip “kwaliteit”, terwijl binnen #4.3 Communication de kernwoorden “communicatie” en “risico” de output bepalen. Opvallen is dat #4.2 Explainability weinig voor komt. Communicatie en herleidbaarheid zijn belangrijk, maar verklaarbaarheid kennelijk minder.

Aan de andere kant van het spectrum zie je dat er een aantal waarden zeer laag scoren. In lijn met de Niveau 1 waarden scoort #5 Diversity, Non-discrimination, and Fairness erg laag. Opvallend is dat het kernbeprip #5.3 Stakeholder Participation helemaal in geen enkel document terugkomt. In de analyse wordt gekeken naar specifieke kernwoorden als “belanghebbenden” of “stakeholders” en die komen in geen enkel document voor.

4.3. Gevoeligheidsanalyse

In de woordanalyse (N3) komen 50 van de 76 woorden uit het model naar voren die in de 400 documenten voorkomen. Van deze woorden zijn er 6 die een impact van meer dan 5% op het totaal hebben.

Tabel 5: Hoogst scorende kernwoorden

Kernwoorden (N3)	Aantal documenten	Aantal keer woord	%totaal
Risico	112	1092	20%
Milieu	60	505	11%
Communicatie	87	741	16%

Privacy	45	287	8%
Kwaliteit	44	434	8%
Control	34	603	6%

Voor deze specifieke kernwoorden is gekeken hoe ze in de teksten terugkomen en of het kernwoord niet te generiek is voor de analyse. Per woord is hieronder aangegeven wat de conclusie is:

Tabel 6: Gevoeligheidsanalyse

Kernwoorden (N3)	Oordeel
Risico	Het woord risico komt in vele varianten in het model voor. Daarbij wordt in de meeste gevallen het woord gebruikt zoals het bedoeld is: "het risico wat men loopt". Vandaar dat besloten is, ondanks het generieke karakter, dit woord wel in de analyse mee te nemen.
Milieu	Het woord milieu komt in een aantal vormen voor. In een aantal gevallen wordt daarbij gesproken over het ministerie van Infrastructuur en Milieu, maar in de meeste gevallen wordt het woord gebruikt in de context zoals die in het model gebruikt wordt (environment). Vandaar dat besloten is dit woord in de analyse op te nemen.
Communicatie	Communicatie is een generiek begrip wat in de specifieke situatie in vele varianten gebruikt wordt[9]. Op basis daarvan is besloten dit woord in de verdere analyse uit te sluiten.
Privacy	Het woord privacy komt in vele varianten voor, maar in de meeste gevallen wel in de betekenis zoals bedoeld in het model. Vandaar dat besloten is dit woord in de analyse op te nemen.
Kwaliteit	Kwaliteit wordt in vele varianten gebruikt. Daarbij worden ook varianten meegenomen (b.v. zorgkwaliteit) die geen relatie tot de bedoelde betekenis hebben. Daarnaast is kwaliteit een redelijk generiek woord. Vandaar dat besloten is deze niet mee te nemen in de analyse.
Control	Control is een woord wat in vele verbasteringen en soorten voor komt. Op basis daarvan is besloten dit woord in de verdere analyse uit te sluiten.
Toezicht	Het woord toezicht komt in vele varianten voor, maar in de meeste gevallen wel in de betekenis zoals bedoeld in het model. Vandaar dat besloten is dit woord in de analyse op te nemen.

Dit resulteert in een aantal kernwoorden die worden uitgesloten van de verdere analyse, te weten "Communicatie", "Kwaliteit" en "Control". In de paragrafen hierna zijn deze woorden uitgesloten.

4.4. Output na gevoeligheidsanalyse

Het uitsluiten van een aantal kernwoorden, veranderd de output zoals eerder weergegeven. In deze paragraaf de veranderde output. Daarbij wordt de vergelijking tussen voor en na gevoeligheidsanalyse in grafische vorm weergegeven in BIJLAGE IV.a.

4.4.1. Niveau 1 output na gevoeligheidsanalyse

Zoals ook in paragraaf 4.2.1 zijn hieronder het aantal documenten weergegeven waarin de verschillende kernwoorden voorkomen.

Tabel 7: N1 resultaten na gevoeligheidsanalyse

Kernwaarde (N1)	Aantal documenten
#1 Human Agency and Oversight	49
#2 Technical Robustness and Safety	46

[9] Telecommunicatie (in meerder varianten), arbeidsmarktcommunicatie, radiocommunicatie, risicocommunicatie, satalietcommunicatie etc…

#3 Privacy and Data governance	95
#4 Transparency	123
#5 Diversity, Non-discrimination, and Fairness	18
#6 Societal and Environmental Well-being	68
#7 Accountability	36

De verschuiving is dat #1 Human Agency and Oversight en #4 Transparency een daling laten zien. Transparantie (vooral op basis van het kernwoord "risico") blijft de belangrijkste kernwaarde. Ook #3 Privacy and Data governance scoort nog steeds boven gemiddeld. De verdeling is niet gelijkmatig.

4.4.2. Niveau 2 output na gevoeligheidsanalyse

Tabel 8: N2 resultaten na gevoeligheidsanalyse

Kernwaarden (N1)	Kernbegrip (N2)	Aantal documenten
#1 Human Agency and Oversight	#1.1 Human Agency and Autonomy	18
	#1.2 Human Oversight	31
#2 Technical Robustness and Safety	#2.1 Resilience to Attack and Security	31
	#2.2 General Safety	8
	#2.3 Accuracy	2
	#2.4 Reliability, Fall-back plans, and Reproducibility	6
#3 Privacy and Data governance	#3.1 Privacy	60
	#3.2 Data governance	35
#4 Transparency	#4.1 Traceability	2
	#4.2 Explainability	7
	#4.3 Communication	114
#5 Diversity, Non-discrimination, and Fairness	#5.1 Avoidance of Unfair Bias	14
	#5.2 Accessibility and Universal Design	4
	#5.3 Stakeholder Participation	0
#6 Societal and Environmental Well-being	#6.1 Environmental Well-being	60
	#6.2 Impact on Work and Skills	3
	#6.3 Impact on Society at large or Democracy	5
#7 Accountability	#7.1 Auditability	2
	#7.2 Risk Management	34

De uitsluiting van de drie kernwoorden heeft ook op Niveau 2 zijn weergave. De waarde van #1.2 Human Oversight daalt door het uitsluiten van het kernwoord "control" van 65 naar 31. Daarnaast daalt de waarde van #4.1 Traceability door het uitsluiten van het woord "kwaliteit" van 77 naar 2 en tenslotte daalt de waarde van #4.3 Communication door het uitsluiten van het woord "communicatie" van 200 naar 114. Daarmee blijft #4.3 Communication de hoogste waarde in de tabel en het belangrijkste kernbegrip. In ondestaande tabel zijn de waarden weergegeven.

4.5. Conclusie

In de empirische analyse is een aggregatie van de kernwoorden naar kernbergrippen (Niveau 2) en kernwaarde (Niveau 1) gemaakt. Daarbij is in de eerste analyse naar voren gekomen dat de waarden voor kernwaarde #4 Transparency met daarbinnen #4.3 Communication de hoogste scores hebben.

Omdat de kernwoorden die gebruikt zijn in een aantal gevallen extreem hoog scoorden is een gevoeligheidsanalyse uitgevoerd. Uit de gevoeligheidsanalyse komt naar voren dat de woorden "Communicatie", "Kwaliteit" en "Control" dusdanig generiek zijn, dat ze uit de analyse zijn uitgesloten. Dit heeft tot gevolg dat de waarde voor #4 Transparency met daarbinnen #4.1 Traceability en #4.3 Communication, maar ook de waarden voor #1 Human Agency and Oversight met daarbinnen #1.2 Human Oversight een daling laten zien. Waarmee de verdeling over de verschilende categoriën nog steeds niet evenredig verdeeld is, maar de verschillen wel kleiner zijn geworden.

Na de empirische analyse zal in hoofdstuk 5 de verbinding tussen alle elementen gemaakt worden en de duiding van de waarnemingen gedaan worden. Daarbij zal alleen gekeken worden naar de implicaties van de metingen na gevoeligheidsanalyse.

5. Discussie, conclusies en aanbevelingen

5.1. Inleiding

De afgelopen hoofdstukken zijn vanuit het theoretisch model, via de methodologie naar het empirisch onderzoek gegaan. In dit hoofdstuk worden de verschillende elementen aan elkaar gekoppeld en daarmee de vraagstelling "Op welke manier kan politiek-strategische governance uitgeoefend worden bij het gebruik van AI in militaire toepassingen?" beantwoord. Daarbij wordt de uitwerking van deelvraag 5 "Welk belang geeft 'de politiek' aan de verschillende meetdimensies?" uitgewerkt op basis van de in het vorige hoofdstuk weergegeven empirische resultaten.

Hoofdstuk 2, het theoretisch model, is geëindigd met twee vervolg-onderzoeksvragen. Deze zullen in paragraaf 5.2 besproken worden. Paragraaf 5.3 zal nog dieper op de duiding van de resultaten ingaan, gevolgd door aanbevelingen voor de praktijk in paragraaf 5.4 en aanbevelingen voor vervolgonderzoek in paragraaf 5.5. Ter afronding worden in paragraaf 5.6 nog de elementen validiteit en betrouwbaarheid behandeld.

5.2. Uitwerking van de vervolg-onderzoeksvragen

Hoofdstuk 2 is afgesloten met een splitsing van de deelvraag 5 ("Welk belang geeft 'de politiek' aan de verschillende meetdimensies?") in een tweetal vervolgvragen:

A. Wordt de governance door de Nederlandse Politiek uitgevoerd met gebruikmaking van de kernwaarden zoals weergegeven in het Trustworthy AI Governance model
B. Is de aanname dat de Nederlandse politiek-strategische governance vooral plaats vindt op basis van de kernelementen wetgeving/monitoring en ethisch/mensenrechten juist?

5.2.1. Gebruik Trusworthy AI Governance model

Kijkende naar de empirische resultaten, dan valt op dat weliswaar alle kernwaarden niveau 1 voorkomen, maar dat de verdeling niet gelijkmatig is. Vooral kernvragen #3 Privancy and Data governance (30%), #4 Transparency (22%) en #6 Societal and Environmental Well-being (16%) komen vaker voor, terwijl kernwaarden als #7 Accountability (8%) en #5 Diversity, Non-discrimination, and Fairness (4%) minder vaak in documenten voor komen.

Dit beeld wordt bevestigd in de analyse op niveau 2. Hierin komen de kernbegrippen #4.3 Communication (26%), #3.1 Privacy en #6.1 Environmental Well-being (beiden 14%) het meeste voor. Terwijl de kernbegrippen uit Niveau 2 voor de Kernwaarde #7 Accountability en #5 Diversity, Non-discrimination, and Fairness weinig voor komen.

Hieruit kan de conclusie getrokken worden dat de deelvraag A "Wordt de governance door de Nederlandse Politiek uitgevoerd met gebruikmaking van de kernwaarden zoals weergegeven in het Trustworthy AI Governance model", positief beantwoord kan worden. Alle kernwaarden uit het model komen terug. Alhoewel niet alle kernwaarden in gelijke hoeveelheid terugkomen in de onderzochte documenten.

5.2.2. Aanname achter het model

Deelvraag B "Is de aanname dat de Nederlandse politiek-strategische governance vooral plaats vindt op basis van de kernelementen wetgeving/monitoring en ethisch/mensenrechten juist?", is lastiger te beantwoorden. Om deze aanname te kunnen toetsen moeten de termen wetgeving/monitoring en ethisch/mensenrechten vertaald worden naar de kernwaarden van het Trustworthy AI Governance model.

Wetgeving/monitoring zou vertaald kunnen worden naar de kernwaarde #3 Privacy and Data governance en #4 Transparency, terwijl ethisch/mensenrechten vertaald zou kunnen worden naar de #5 Diversity, Non-discrimination, and Fairness en #6 Societal and Environmental Well-being. Bij elkaar geteld vertegenwoordigen deze kernwaarden 70% van de voorkomende waarden.

Het model bestaat uit 7 kernwaarden. Bij een gelijke verdeling zouden 4 kernwaarden gezamenlijk 57% van het model vertegenwoordigen. De eerder vermelde 4 kernwaarden hebben gezamenlijk 70%.

Dit in overweging nemende acht ik de aanname dat de governance vooral op wetgeving/monitoring en ethisch/mensenrechten plaats vindt, onderbouwd. De afwijking tov de normaal-verdeling is significant. Aandachtspunt is wel dat ook de laagste waarde, #5 Diversity, Non-discrimination, and Fairness, in de selectie zit. Dit is echter onvoldoende om daarmee de hele aanname ongeldig te verklaren.

5.3. Interpretatie resultaten

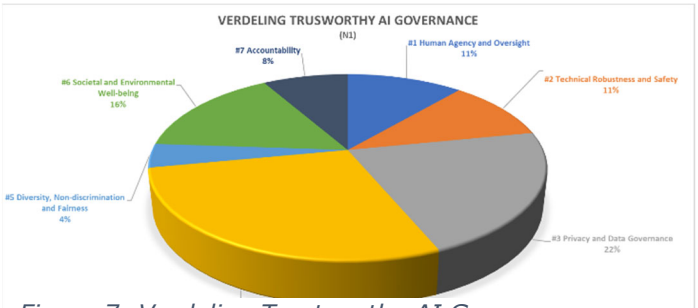
Tot nu toe was de analyse steeds vraag en antwoord gedreven. In deze paragraaf wordt ingegaan op de samenhang van de vragen en antwoorden. Wat betekent dit antwoord? Hoe moet het antwoord geïnterpreteerd worden? Daarbij worden een tweetal verschillende dimensies bekeken, te weten de relatie tussen de literatuur en het empirisch onderzoek (par 5.3.1.) en de interpretatie van de resultaten (par 5.3.2).

5.3.1. Relatie tussen empirische resultaten en literatuur

Trustworthy AI is een relatief nieuwe methodiek. De eerste versie van het model is op 17 juni 2020 gepubliceerd. Als gevolg daarvan zijn er, voor zover op dit moment bekend, nog geen empirische onderzoeken rondom de theorie gedaan. Dit onderzoek is daarmee een van de eerste toetsingen van de theorie aan de praktijk.

In het theoretische model zie je dat alle 7 kernwaarden ongeveer even veel aandacht krijgen. De beschrijving van iedere waarde en de onderliggende detailvragen variëren beperkt in omvang. Iedere waarde wordt in 2 tot 2 ½ bladzijde beschreven. Hieruit zou je af kunnen leiden dat er niet één kernwaarde is die veel belangrijker is dan de andere.

De verwachting is dan, dat de empirische resultaten ook een gelijkmatige verdeling laten zien. In het empirisch onderzoek zie je dit niet als zodanig terugkomen. De aandacht van de Nederlandse politiek is dus niet gelijkmatig over de kernwaarde verdeeld. In figuur 7 is dit grafisch weergegeven.



Figuur 7: Verdeling Trustworthy AI Governance kernwaarden

#4 Transparency (28%) en #3 Privacy and Data governance (22%) omvatten samen de helft van alle metingen, terwijl de kernwaarden #5 Diversity, Non-discrimination, and Fairness (4%) en #7 Accountability (8%) veel minder aandacht krijgen. De overige drie kernwaarden krijgen ongeveer dezelfde aandacht.

Conclusie hieruit zou zijn dat, waar het model uit gaat van redelijk gelijkmatige verdeling, de focus in de praktijk naar een beperkt aantal kernwaarden zal gaan. Politieke voorkeuren, specifieke situationele gebeurtenissen of modelmatige onvolkomenheden zouden hier de oorzaak van kunnen zijn, maar dit is niet uit de data af te leiden. Hiervoor zou aanvullend onderzoek noodzakelijk zijn.

5.3.2. Interpretatie van de resultaten

De ongelijkmatige verdeling over de kernwaarden trekt zich in de details door naar kernbegrippen en kernwoorden. Als je naar Niveau 2 gaat, zie je dat het kernbegrip #5.3 Stakeholder Participation in geen enkel document terugkomt. De woorden “belanghebbende” en “stakeholder” komen in geen enkel document voor. Daarmee kun je echter niet de conclusie trekken dat er niet over stakeholders gesproken wordt. Alleen worden deze niet in die hoedanigheid benoemd. Het woord “slachtoffer” en “medewerker” komt ook voor. Dat zijn ook belanghebbenden. Echter, ze worden in de documenten niet gebruikt in de zin van betrokkenheid van belanghebbenden. Een extra filter op het woord “betrokkenheid” levert wel twee relevante referenties op (bijvoorbeeld

*"We zien een toenemende **betrokkenheid** van burgers"*[10]). Maar ook dat is nog steeds een lage score.

Hieruit zouden een tweetal voorzichtige conclusies getrokken mogen worden: Als eerste dat er maar beperkt over stakeholders gesproken wordt. Zelfs als er andere zoekwoorden gebruikt worden, blijft de score laag. Daarnaast zou de constatering kunnen zijn dat, als er over stakeholders gesproken wordt, dit gebeurt in indirecte bewoordingen. Echter dat laatste is met de gekozen methodiek niet te meten.

Ook op niveau 3 zie je een aantal kernwoorden in geen enkel document terugkomen: "accuraat", "traceerbaarheid", "AVG", "handicap", "herleidbaar" om maar een paar voorbeelden te noemen. Woorden die gevoelsmatig toch redelijk gangbaar zijn in het taalgebruik en die je wel zou verwachten. Ook hierbij kan de keuze van de kernwoorden een factor van invloed zijn. Als bijvoorbeeld het woord "nauwkeurig" gebruikt zou worden i.p.v. "accuraat". Dan zou dat één referentie opleveren (bijvoorbeeld *"om aan te geven dat een krijgsmacht vele **nauwkeurige** stappen doorloopt, voordat over wordt gegaan tot het uitvoeren van een (aanvals)missie"*[11]). Ook hier zou de constatering zijn dat deze wijziging van kernwoord geen drastische wijziging in de meting tot gevolg heeft.

Ook waar de score hoog is, zoals bijvoorbeeld bij #4 Transparency, zie je dat dit veroorzaakt wordt door een beperkt aantal kernwoorden. Het kernwoord "risico" komt in 112 documenten voor, wat leidt tot een hoge score voor het kernbegrip #4.3 Communication (114) wat wederom leidt tot de hoge score op #4 Transparency (123). Vooral hierbij geld dat de keuze van het woord een grote impact op de meting heeft. In de gevoeligheidsanalyse is daarom specifiek naar de woorden met een hoge score gekeken en is bewust besloten om het woord risico in de analyse te laten. Risico is een onderwerp waar vaak over gesproken wordt. Ook in de context van AI (bijvoorbeeld *"Ten aanzien van de inzet van autonome wapensystemen zijn guiding principles aangenomen. De leden van de VVD-fractie vragen hoe de regering omgaat met de mogelijkheid van artificial intelligence in de responsiviteit. Door het «human-out-of-the-loop» principe is er **risico** van een «hyperwar»"*[12])

Kortom, er zijn dus kernwoorden die meer en kernwoorden die minder gebuikt worden. Woorden als "toezicht", "privacy", "integriteit", "risico" en "milieu" zijn veel voorkomende woorden. Er wordt dus meer over de elementen #1.2 Human Oversight, #3 Privacy and Data governance, #4 Transparency, #4.3 Communication en #6.1 Environmental Well-being gesproken.

5.4. Aanbevelingen voor de praktijk

Het feit dat er bepaalde kernwoorden, kernbegrippen en kernwaarden zijn die vaker gebruikt worden dan andere heeft implicaties voor de praktijk. In de praktijk zal hier mee omgegaan moeten worden. Er ontstaat een belang van volledigheid en focus. Focus op de onderwerpen die aandacht krijgen, maar ook bewaken van de volledigheid van onderwerpen die behandeld worden.

In de huidige politiek-strategische governance worden bepaalde kernwoorden/-begrippen/-waarden vaker gebruikt. De focus ligt meer op transparantie, data governance en privacy. Mogelijk zijn dat, op dat moment de elementen waarover gesproken moet worden. Maar het risico bestaat ook dat er elementen vergeten worden. Voor de praktijk, lees politici als uitvoerende personen en Bestuursstaf, binnen het Ministerie van Defensie, als ondersteunend element, ligt hier een verantwoordelijkheid om de volledigheid te bewaken. Alle elementen die een rol spelen in het model zullen behandeld moeten worden, waarbij de juiste balans gezocht wordt.

Voor de betrokkenen houdt dit een aantal zaken in. Allereerst zullen alle betrokkenen bekend moeten zijn met het model. Welke kernwaarden spelen een rol binnen de governance van AI en hoe worden die toegepast. De betrokken medewerkers van de Bestuursstaf zullen bekend moeten zijn met het model, maar deze ook uit moeten dragen richting de politici.

[10] 2014D25665: Trends transities TNO
[11] 2021D20388: Initiatiefnota van het lid Belhaj over Autonome Wapensystemen
[12] 2019D33145: Verslag van een schriftelijk overleg over de geannoteerde agenda informele Raad Buitenlandse Zaken Defensie van 28-29 augustus 2019 te Helsinki

Aan de andere kant is de aandacht voor bepaalde onderwerpen groter. Voor de Bestuursstaf is het dan aan te raden dat er, bij de voorbereiding van documenten, extra aandacht gegeven wordt aan de onderwerpen transparantie, data goverance en privacy. Door deze onderwerpen extra uit te diepen, kunnen vragen vermeden worden, of op voorhand beantwoord worden. Hierdoor kan de politieke discussie meer evenredig verdeeld worden, zodat ook naar de andere kernwaarden gekeken wordt.

5.5. Aanbevelingen voor vervolgonderzoek

Zoals alle onderzoeken, is ook dit onderzoek begrensd in tijd en ruimte. In deze begrenzing liggen dan ook de aanbevelingen voor vervolgonderzoek.

Allereerst is het onderzoek begrensd in tijd. Het onderzoek is gedaan op een dataset van 10 jaar (maart 2011- juni 2021). Voor deze periode is gekozen omdat de eerste documenten over kunstmatige intelligentie uit maart 2011 stammen. In deze periode was de politieke situatie in Nederland redelijk stabiel. Alhoewel er in de onderzochte periode drie kabinetten zijn geweest, was de samenstelling van de regering redelijk constant qua politieke voorkeur (centrum rechts). Dit zou tot gevolg kunnen hebben dat debatten die gevoerd worden en retoriek die gebruikt wordt, ook redelijk constant zijn. Het zou daarom aan te bevelen zijn dit onderzoek op een later tijdstip nogmaals te doen als de samenstelling van de regering is veranderd, of als aanvullende informatie over politieke voorkeur beschikbaar is.

Het onderzoek is ook begrensd tot de Nederlandse politieke situatie. De resultaten van het onderzoek hoeven daarmee nog niet valide te zijn voor andere landen binnen de EU of binnen de wereld. Hiervoor is aanvullend onderzoek nodig.

Dit onderzoek is een kwantitatief onderzoek. In een kwantitatief content onderzoek wordt een hypothese getoetst. Kwalitatieve analyses zijn vooral gericht op hypothese vorming. De keuze voor de hypothese van de Trustworthy AI Governance is gedaan op basis van de aanname dat de Nederlandse politiek-strategische governance vooral plaats vindt op basis van de kernelementen wetgeving/monitoring en ethisch/mensenrechten. Deze aanname is getoetst. De kwantitatieve analyse van het Trustworthy AI Governance model geeft bruikbare resultaten. Daarbij rijst wel de vraag of er andere modellen zijn die een hogere verklarende waarde kunnen hebben. Dit zou onderzocht kunnen worden door andere modellen kwantitatief te onderzoeken op dezelfde dataset of door een kwalitatieve analyse van de meest relevante documenten binnen de beschikbare brondata.

Een specifieke afwijking die terugkomt is het kernbegrip “Stakeholder”. In de analyse kwam naar voren dat dit kernbegrip in geen enkel document terugkwam. Dat zou kunnen komen omdat er niet over belanghebbenden wordt gesproken, of doordat er alleen indirect over belanghebbenden wordt gesproken. Uit een snelle (kwalitatieve) analyse blijkt het laatste het geval te zijn, maar aanvullende kwalitatieve analyse is noodzakelijk om dit definitief vast te stellen.

5.6. Betrouwbaarheid en validiteit

Zoals in hoofdstuk 3.5 aangegeven is de betrouwbaarheid van een wordscore onderzoek hoog, Maar zijn er ook zaken die de betrouwbaarheid negatief beïnvloeden:

- Handmatige conversies: Doordat er veel handmatige slagen nodig waren voor conversie zijn er veel controles uitgevoerd. Deze hebben tijdens de verwerking tot correcties geleid om fouten te herstellen. Achteraf is alleen niet vast te stellen of en hoeveel afwijkingen niet gevonden zijn.
- Gelijkwaardig lexicografisch niveau: Omdat het allemaal teksten betreft van de discussies en behandelingen van de Tweede Kamer, is de aanname dat aan deze voorwaarde is voldaan.

Daarnaast is in hoofdstuk 3.5 ook aangegeven dat validiteit een punt van aandacht is voor wordscore-onderzoek. Een aantal van de (veel voorkomende) woorden is in de tekst opgezocht om de relatie tussen het onderwerp, de governance van AI en het betreffende woord vast te stellen. In de voorgaande paragraaf, maar ook in de hieronder staande Tabel 9: Citaten met kernwoorden in

de documenten, zijn een aantal voorbeelden van de gebruikte kernwoorden in de context weergegeven.

Tabel 9: Citaten met kernwoorden in de documenten

"*overwegende dat er ook serieuze risico's kleven aan het gebruik van big data, algoritmen en analysemethoden, met discriminatie of schending van **privacy** of andere grondwetten als potentieel gevolg*" (n.a.v. debat over Big Data en Kunstmatige Intelligentie)[13]
"*Menselijk **toezicht** en mogelijkheden tot ingrijpen*" (een van de zeven elementen bij inzet autonome wapensystemen)[14]
*"De leden van de GroenLinks-fractie vragen de regering om toe te lichten hoe bij het standpunt van de regering om tegen een verbod of moratorium op de ontwikkeling van autonome wapens te zijn is meegewogen dat autonome wapens het **risico** op (het escaleren van) conflicten is meegewogen."*[15]
*"Met Europese collegae en deskundigen is gesproken over o.a. de **risico**'s van kunstmatige intelligentie, met name met betrekking tot de ontwikkeling van autonome wapens, en mogelijke regelgeving hiervoor"*[16]

[13] Plenair verslag Tweede Kamer, 90e vergadering, woensdag 6 juni 2018
[14] 2021D20388: Initiatiefnota van het lid Belhaj over Autonome Wapensystemen
[15] 2019D33145: Verslag van een schriftelijk overleg over de geannoteerde agenda informele Raad Buitenlandse Zaken Defensie van 28-29 augustus 2019 te Helsinki
[16] 2019D19128: Internationaal normkader voor het gebruik van nieuwe technologieën als (onderdeel van) wapensystemen

BIJLAGE I. Literatuuronderzoek resultaten

In onderstaande tabel zijn de zoekresultaten weergegeven. Daarbij wordt per deelvraag aangegeven welke zoektermen in welke bron gebruikt zijn, met welke filters en welke sortering. De sortering is vooral van belang omdat bij de sortering "relevance" gebruik wordt gemaakt van het algoritme van de site en deze op een later tijdstip, of door anderen, niet reproduceerbaar is.

Binnen de tabel worden een aantal kolommen gebruikt om weer te geven wat de resultaten van de zoekslagen zijn. Daarbij worden de volgende kolommen gebruikt:

- Resultaat: Het aantal wat totaal uit de zoekslag naar voren komen.
- Bekeken: Dit zijn de uitkomsten van de zoektocht waarbij de titel en de summary bekeken en beoordeeld zijn.
- Geselecteerd: Vanuit de bekeken artikelen zijn een aantal interessant genoeg om in detail te bekijken. De interessante objecten zijn in Endnote opgeslagen als refference en indien beschikbaar ook als PDF/MP4 aan de referentie toegevoegd. In de kolom is het totaal aantal weergegeven.

Tabel 10: Overzicht zoekresultaten literatuuronderzoek in detail deelvragen 1-4

Deelvraag	Zoektermen	Resultaten	Bekeken	Geselecteerd
1. Wat is AI in een militaire context	Open bronnen/referenties collega's	56	56	7
	Zoektermen: (Artificial intelligence) AND (definition) AND (Military) Bron: OU Bibliotheek Periode filter 01-01-2018 – 31-12-2020 Termen onderwerp: "Artificial intelligence"	358	358	5
	Zoektermen: "Artificial intelligence" AND "military" AND "applications" Bron: Google Scholar Periode filter sinds 2020 Sortering: relevance	8760	30[17]	6
	Zoektermen: (stages of Aritificial intelligence) Bron: OU Bibliotheek	33	33	1
	Zoektermen: XAI Bron: OU Bibliotheek Periode filter: 2019-2020 Onderwerp: artificial intelligence	56	56	3
	Zoektermen: (XAI) AND (model) AND (Stages) Bron: Google Scholar Periode filter: sinds 2019 Sortering: relevance	1109	120[18]	2
	Research Gate Zoekterm "Military Applications of artificial intelligence"	30+	10[19]	1

[17] Door de sortering relevance worden de volgens het algoritme meest relevance artikelen vooraan geplaatst. De artikelen zijn daarna beoordeeld (titel en samenvatting) totdat de relevance zichtbaar afneemt. Daarna is het zoekproces beëindigt.
[18] Door de sortering relevance worden de volgens het algoritme meest relevance artikelen vooraan geplaatst. De artikelen zijn daarna beoordeeld (titel en samenvatting) totdat de relevance zichtbaar afneemt. Daarna is het zoekproces beëindigt.
[19] Door de sortering relevance worden de volgens het algoritme meest relevance artikelen vooraan geplaatst. De artikelen zijn daarna beoordeeld (titel en samenvatting) totdat de relevance zichtbaar afneemt. Daarna is het zoekproces beëindigt.

2. Hoe kan governance van AI worden ingericht	Zoektermen: (Artificial intelligence) AND (Governance) AND (Military) AND (Model) Bron: OU Bibliotheek Periode filter: 01-01-2018 – 31-12-2020 Termen onderwerp: "Artificial intelligence"	121	121	6
	Zoektermen: governance models "Artificial intelligence" military Bron: Google Scholar Periode filter: 2020 Sortering: relevence	1390	120[20]	3

[20] Door de sortering relevance worden de volgens het algoritme meest relevance artikelen vooraan geplaatst. De artikelen zijn daarna beoordeeld (titel en samenvatting) totdat de relevance zichtbaar afneemt. Daarna is het zoekproces beëindigt.

BIJLAGE II. Vertaling Trustworthy Goverance in steekwoorden

Tabel 11: Vertaling Trustworthy governance in steekwoorden

Kernwaarde (N1)	Kernbegrip (N2)	Kernwoorden (ENG)	kernwoorden (N3)
#1 Human Agency and Oversight	Human Agency and Autonomy	· Human end-user	Menselijk
		· Confusion for end-users	Verwarring
		· User awareness of outcome AI	Bewustzijn
		· User awareness working with AI	
		· To-much reliance on AI	Sociale
		· Social interaction	Interactie
		· Negative consequences	Consequenties
		· Dependability	Afhankelijk
		· Risk of addiction	Verslaving
		· Risk of manipulation	Manipulatie
	Human Oversight	· Human-in-the-loop	
		· Human-on-the-loop	Beheersing
		· Human-in-control	Control
		· Training on oversight	Toezicht
		· Detection and response mechanism	
		· Stop button	Zelflerend
		· Oversight and control self-learning	Leren
#2 Technical Robustness and Safety	Resilience to Attack and Security	· Adversarial, critical or damaging effects	Schadelijk
		· Cybersecurity, -attacks	Cybersecurity
		· Vulnerabilities, entry points	
		· Data poisoning	Gegevensvergiftiging
		· Model evasion	
		· Model inversion	Integriteit
		· Integrity, robustness, overall security	Weerbaar
		· Red-team, pentest	Resilience
		· Security coverage	Security
		· Security updates	Updates
	General Safety	· Risks, risks metrics and risks levels	Risicoindicatoren
		· Continuous measure and assess risks	Risiconiveau's
		· Malicious use, misuse of inappropriate use	Misbruik
		· Critical safety levels	Safety
		· Fault tolerance	Fouttolerantie
		· Technical robustness and safety	Robuust
	Accuracy	· Critical, adversarial of damaging consequences	Vijandig
		· Measure to ensure correct data	Juistheid
		· Up-to-date, high quality, complete and representative	
		· Monitor and document accuracy	Acuraat
		· Validate the data it was trained on	

		· Accuracy properly communicated	
	Reliability, Fall-back plans, and Reproducibility	· Critical, adversarial or damaging consequences	Vijandig
		· Human safety	Betrouwbaar
		· Test specific contexts	
		· Reproducibility	Reproduceerbaar
		· Verification and validation methods and documentation	Herhaalbaar
		· Document operational process	
		· Failsafe fallback	
		· Low confidence score	
		· Potential negative consequences	Gevolgen
#3 Privacy and Data governance	Privacy	· Impact on privacy	Privacy
		· Data protection	Gegevensbescherming
		· Physical, mental, and moral integrity	Integriteit
		· Issues related to privacy	
	Data governance	· General Data Protection Regulation (GDPR)	GDPR
		· Data Protection Officer (DPO)	AVG
		· Oversight mechanisms	Oversight
		· Privacy-by-design	Toezicht
		· Encryption, pseudonymization, aggregation, anonymization	Encriptie
		· Withdraw consent	
		· Privacy and data implications	
#4 Transparency	Traceability	· Traceability	Traceerbaarheid
		· Assess quality	Kwaliteit
		· Trace back data	
		· Trace back model or rules	
		· Logging	Logging
	Explainability	· Data minimization	Minimalisatie
		· Explain decisions	Toelichten
		· Survey users understanding	Begrip
	Communication	· Inform users	Informeren
		· Communicate benefits	Communicatie
			Voordelen
		· Technical limitations	Beperkingen
		· Potential risks	Risico
#5 Diversity, Non-discrimination, and Fairness	Avoidance of Unfair Bias	· Avoid creating or reinforcing unfair bias	Vooringenomenheid
		· Diversity and representation	Diversiteit
		· Understanding data, model and performance	
		· Education and awareness initiatives	Bewustzijn
		· Flag of issues	Signaleren
		· Indirect effects	
		· Fairness	Eerlijk
		· Impacted communities	Gemeenschap

		· Quantitative analysis of metrics to measure	
	Accessibility and Universal Design	· Preferences and abilities	Toegankelijkheid
		· Useable by those with special needs or disabilities	Handicaps
		· Consult end-users	
		· Impact of the AI-system on users	
		· Groups disproportionately affected	Evenredig
	Stakeholder Participation	· Participation of the stakeholders during design and development	Belanghebbenden
#6 Societal and Environmental Well-being	Environmental Well-being	· Negative impact environment	Milieu
		· Evaluate impact	Evalueren
		· Reduce impact	
	Impact on Work and Skills	· Impact human work and work arrangements	
		· De-skilling the workforce	Medewerkers
		· Use of new (digital) skills	Personeelsleden
	Impact on Society at large or Democracy	· Indirectly affected stakeholders	Belanghebbenden
		· Potential harm	Leed
		· Negative impact democracy	Democratie
#7 Accountability	Auditability	· Auditability	Controleerbaarheid
		· Traceability	Herleidbaar
		· By independent third parties	Onafhankelijk
	Risk Management	· Ethical concerns	Etisch
		· Accountability measures	Verantwoording
		· Legal framework	Juridisch
		· Ethics review board	
		· Identification and documentation of conflicts	Verslaglegging
		· Appropriate training	Training
		· Report potential vulnerabilities	Gevoeligheden
		· Revision of the risk management process	
		· Adversely affect individuals	
		· Mechanism in place	

BIJLAGE III. Uitwerken van de analyse in detail

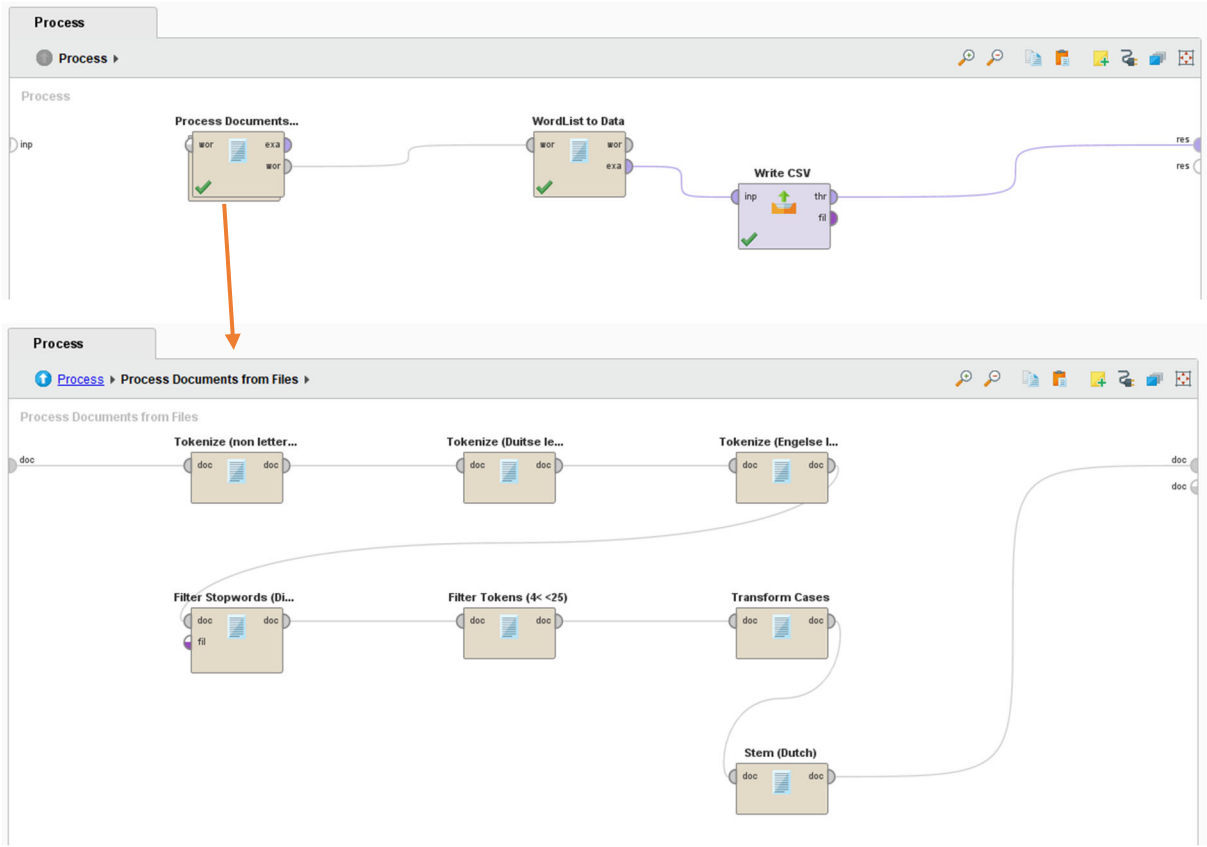
a. Selecteren van de bestanden

De documenten zijn handmatig geselecteerd en gefilterd. Daarbij is gebruik gemaakt van de windows indexeerfunctie. Om de uiteindelijke selectie te maken zijn de verschillende termen uit de categorisering gebruikt. Uiteindelijk hebben alleen onderstaande vier kernbegrippen tot selectie van documenten geleid. De overige woorden leverden geen extra hits op:

1) "kunstmatige intelligentie"
2) "Artificial intelligence"
3) "Robotisering"
4) "Autonome wapen"

Dit levert totaal 400 bruikbare documenten op.

b. Verwerken van de bestanden tot word-score



Process documents from files (subproces)

- Tokenize (non-letters)
- Tokenize (Duits)
- Tokenize (Engels)

Met de functie Tokenize worden specifieke leestekens geëlimineerd. Denk daarbij aan β£≤√Ω

- Filter Stopwords (Dutch)

In een taal zitten vaak woorden die voor het spreken van de taal een verbindende werking hebben, maar geen verklarende waarde. Denk bijvoorbeeld aan "de, het, een, die, dat" etc. Op basis van een specifieke Nederlandse stopwoordenlijst[21] zijn deze woorden eruit gefilterd.

- Filter Tokens by length (4< >25)

Om de woordelnijst te reduceren zijn de zeer kleine woorden en de zeer lange woorden eruit gefilterd.

- Transform cases to lower cases

Omdat de software hoofd en kleine letters als verschillende tekens ziet, zijn alle woorden omgezet in kleine letters.

- Stemming (Dutch)

Met de functie stemming worden woorden omgezet in de stam-vorm.

- Wordlist from data

Met deze functie wordt de tekst daadwerkelijk in een wordlijst omgezet.

- Write to csv

De output is als csv opgeslagen op de schijf om daarna in excel verwerkt te kunnen worden.

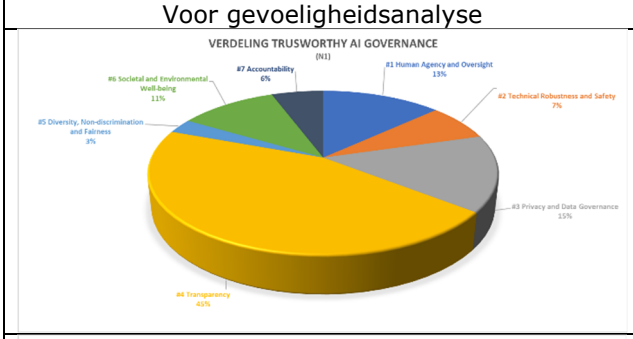
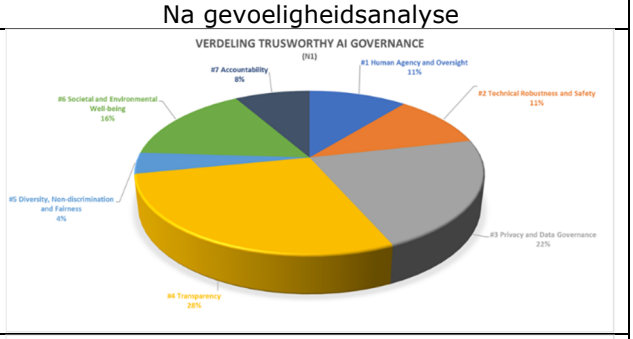
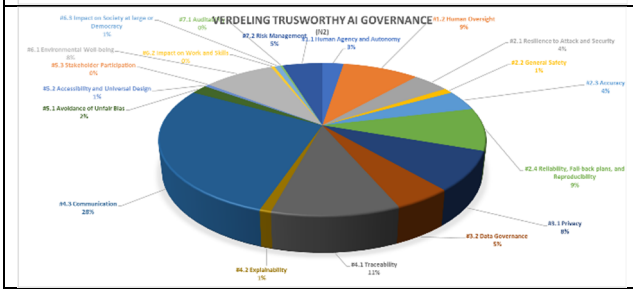
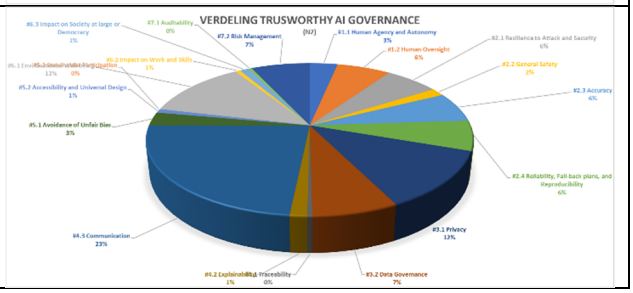
De matching met de sleutelwoorden heeft uiteindelijk in Excel plaatsgevonden. De systeemtechnische belasting van Rapidminer was zo groot dat dit met de beschikbare machines niet mogelijk was.

[21] https://countwordsfree.com/stopwords/dutch

BIJLAGE IV. resultaten

a. Resultaten voor en na gevoeligheidsanalyse

In onderstaande grafieken zijn de resultaten weergegeven voor en na de gevoeligheidsanalyse

Voor gevoeligheidsanalyse	Na gevoeligheidsanalyse
	
	

b. Totaaloverzicht Resultaten voor en na gevoeligheidsanalyse

Kernwaarde (N1)	#DOC N1 vo	#DOC N1 n	Kernbegrip (N2)	#DOC N2 voo	#DOC N2 na2	Kernwoorden (ENG)	kernwoorden (N3)	#DOC N3	#KEER N3
#1 Human Agency and Oversight	83	49	Human Agency and Autonomy	18	18	• Human end-user	Menselijk	2	192
						• Confusion for end-users	Verwarring	0	0
						• User awareness of outcome AI	Bewustzijn	1	61
						• User awareness working with AI		0	0
						• To-much reliance on AI	Sociale	0	0
						• Social interaction	Interactie	6	130
						• Negative consequences	Consequenties	1	173
						• Dependability	Afhankelijk	4	21
						• Risk of addiction	Verslaving	2	7
						• Risk of manipulation	Manipulatie	2	23
			Human Oversight	65	31	• Human-in-the-loop		0	0
						• Human-on-the-loop	Beheersing	1	2
						• Human-in-control	~~control~~	~~34~~	~~693~~
						• Training on oversight	Toezicht	27	350
						• Detection and response mechanism		0	0
						• Stop button	Zelflerend	1	2
						• Oversight and control self-learning	Leren	2	54
#2 Technical Robustness and Safety	46	46	Resilience to Attack and Security	31	31	• Adversarial, critical or damaging effects	Schadelijk	2	60
						• Cybersecurity, -attacks	Cybersecurity	8	126
						• Vulnerabilities, entry points		0	0
						• Data poisoning	Gegevensvergiftiging	0	0
						• Model evasion		0	0
						• Model inversion	Integriteit	14	137
						• Integrity, robustness, overall security	Weerbaar	2	23
						• Red-team, pentest	Resilience	0	0
						• Security coverage	Security	4	154
						• Security updates	Updates	1	34
			General Safety	8	8	• Risks, risks metrics and risks levels	Risicoindicatoren	0	0
						• Continuous measure and assess risks	Risiconiveau's	0	0
						• Malicious use, misuse of inappropriate use	Misbruik	5	149
						• Critical safety levels	Safety	1	29
						• Fault tolerance	Fouttolerantie	0	0
						• Technical robustness and safety	Robuust	2	121
			Accuracy	2	2	• Critical, adversarial of damaging consequences	Vijandig	1	2
						• Measure to ensure correct data	Juistheid	1	17
						• Up-to-date, high quality, complete and representative		0	0
						• Monitor and document accuracy	Acuraat	0	0
						• Validate the data it was trained on		0	0
						• Accuracy properly communicated		0	0
			Reliability, Fall-back plans, and Reproducibility	6	6	• Critical, adversarial or damaging consequences	Vijandig	1	2
						• Human safety	Betrouwbaar	4	25
						• Test specific contexts		0	0
						• Reproducibility	Reproduceerbaar	0	0
						• Verification and validation methods and documentation	Herhaalbaar	0	0
						• Document operational process		0	0
						• Failsafe fallback		0	0
						• Low confidence score		0	0
						• Potential negative consequences	Gevolgen	1	4
#3 Privacy and Data Governance	95	95	Privacy	60	60	• Impact on privacy	Privacy	45	287
						• Data protection	Gegevensbescherming	1	3
						• Physical, mental, and moral integrity	Integriteit	14	137
						• Issues related to privacy		0	0
			Data Governance	35	35	• General Data Protection Regulation (GDPR)	GDPR	1	8
						• Data Protection Officer (DPO)	AVG	0	0
						• Oversight mechanisms	Oversight	2	10
						• Privacy-by-design	Toezicht	32	362
						• Encryption, pseudonymization, aggregation, anonymization	Encriptie	0	0
						• Withdraw consent		0	0
						• Privacy and data implications		0	0
#4 Transparency	284	123	Traceability	77	2	• Traceability	Traceerbaarheid	0	0
						• Assess quality	~~Kwaliteit~~	~~75~~	~~625~~
						• Trace back data		0	0
						• Trace back model or rules		0	0
						• Logging	Logging	2	6
			Explainability	7	7	• Data minimization	Minimalisatie	1	3
						• Explain decisions	Toelichten	0	0
						• Survey users understanding	Begrip	6	224
			Communication	200	114	• Inform users	Informeren	1	3
						• Communicate benefits	~~Communicatie~~	~~87~~	~~741~~
							Voordelen	0	0
						• Technical limitations	Beperkingen	1	2
						• Potential risks	Risico	112	1092
#5 Diversity, Non-discrimination and Fairness	18	18	Avoidance of Unfair Bias	14	14	• Avoid creating or reinforcing unfair bias	Vooringenomenheid	0	0
						• Diversity and representation	Diversiteit	4	137
						• Understanding data, model and performance		0	0
						• Education and awareness initiatives	Bewustzijn	1	61
						• Flag of issues	Signaleren	0	0
						• Indirect effects		0	0
						• Fairness	Eerlijk	3	184
						• Impacted communities	Gemeenschap	6	281
								0	0
						• Quantitative analysis of metrics to measure		0	0
			Accessibility and Universal Design	4	4	• Preferences and abilities	Toegankelijkheid	3	6
						• Useable by those with special needs or disabilities	Handicaps	0	0
						• Consult end-users		0	0
						• Impact of the AI-system on users		0	0
						• Groups disproportionately affected	Evenredig	1	4
			Stakeholder Participation	0	0	• Participation of the stakeholders during design and development	Belanghebbenden	0	0
#6 Societal and Environmental Well-being	68	68	Environmental Well-being		60	• Negative impact environment	Milieu	60	505
						• Evaluate impact	Evalueren	0	0
						• Reduce impact		0	0
			Impact on Work and Skills	3	3	• Impact human work and work arrangements		0	0
						• De-skilling the workforce	Medewerkers	3	9
						• Use of new (digital) skills	Personeelsleden	0	0
			Impact on Society at large or Democracy	5	5	• Indirectly affected stakeholders	Belanghebbenden	0	0
						• Potential harm	Leed	0	0
						• Negative impact democracy	Democratie	5	131
#7 Accountability	36	36	Auditability	2	2	• Auditability	Controleerbaarheid	0	0
						• Traceability	Herleidbaar	0	0
						• By independent third parties	Onafhankelijk	2	14
			Risk Management	34	34	• Ethical concerns	Etisch	0	0
						• Accountability measures	Verantwoording	11	76
						• Legal framework	Juridisch	1	231
						• Ethics review board		0	0
						• Identification and documentation of conflicts	Verslaglegging	2	4
						• Appropriate training	Training	20	214
						• Report potential vulnerabilities	Gevoeligheden	0	0
						• Revision of the risk management process		0	0
						• Adversely affect individuals		0	0
						• Mechanism in place		0	0

BIJLAGE V. Referenties

Allen, G., & Chan, T. (2017). *Artificial Intelligence and National Security - A study on behalf of Dr. Jason Matheny, Director of the U.S. Intelligence Advanced Research Projects Activity (IARPA)*. Cambrdicge MA: HARVARD Kennedy School BELFER CENTER for Science and International Affairs.

Almeida, P., Santos, C., & Farias, J. S. (2020). *Artificial Intelligence Regulation: A Meta-Framework for Formulation and Governance.* Paper presented at the Proceedings of the 53rd Hawaii International Conference on System Sciences.

Benoit, W. L. (2014). Content analysis in political communication. In *Sourcebook for political communication research* (pp. 290-302): Routledge.

Bogdanoski, M., & Nacev, A. THE USE OF ARTIFICIAL INTELLIGENCE IN THE MILITARY DOMAIN.

Corn, J. D. (2019). *DoD Artificial Intelligence Strategy Overview*. Retrieved from

Cotter, T. S. (2015). RESEARCH AGENDA INTO HUMAN-INTELLIGENCE/MACHINE-INTELLIGENCE GOVERNANCE. *Proceedings of the International Annual Conference of the American Society for Engineering Management*, 1.

(2019). *AI Strategy, Policy, and Governance* [Retrieved from https://www.youtube.com/watch?v=2IpJ8TIKKtI&t=337s

Defensiestaf. (2019). *Nederlandse Defensie Doctrine*. Den Haag: Defensiestraf

Ducheine, P. A. L. (2008). *Krijgsmacht, Geweldgebruik & Terreurbestrijding*: Wolf Legal Publishers.

Gasser, U., & Almeida, V. A. F. (2017). A Layered Model for AI Governance. *IEEE Internet Computing, 21*(6), 58-62. doi:10.1109/mic.2017.4180835

Grega, M., SabÓ, A., NeČAs, P., Simulation Centre, A. F. A. i. L. M. D. L. M. S. a. s. a. s., Simulation Centre, A. F. A. i. L. M. D. L. M. S. m. g. a. s., Department of Security Studies, F. o. P. S., & International Relations, M. B. U. i. B. B. K. B. B. S. p. n. u. s. (2019). AI in Military Synthetic Simulation Environment of the Slovak Republic. In (Vol. 11, pp. 211-219): National Institute for Aerospace Research “Elie Carafoli” - INCAS.

(2019). *Keep your AI under Control - Governance of AI* [

High-Level_Expert_Group_on_Artificial_Intelligence. (2020). *Assessment List for Trustworthy Artificial Intelligence*. Retrieved from https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342

Hoadley, D. S., & Sayler, K. M. (2019). *Artificial Intelligence and National Security*.

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *BUSINESS HORIZONS, 62*(1), 15-25. doi:10.1016/j.bushor.2018.08.004

Laver, M., Benoit, K., & Garry, J. (2003). Extracting Policy Positions from Political Texts Using Words as Data. *The American political science review, 97*(2), 311-331. doi:10.1017/S0003055403000698

Manen, H. v., Sweijs, T., & Arkhipov-Goyal, A. (2019). *Macro Implications of Micro Transformations: An Assessment of AI’s Impact on Contemporary Geopolitics*. Retrieved from Den HAag:

Morgan, F. E., Boudreaux, B., Lohn, A. J., Ashby, M., Curriden, C., Klima, K., & Grossman, D. (2020). *Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World*. Retrieved from

Nefes, S. T. (2020). *Using Content Analysis to Study Political Texts: Notes on Turkish Parliamentary Debates.* Spanish National Research Council (CSIC), Madrid, Spain.

Siau, K., & Wang, W. (2018). *Artificial Intelligence: A Study on Governance, Policies, and Regulations*.

Voetelink, J. E. D. (2012). Een introductie in het militair operationeel recht. *Militaire Spectator, 181-1*.